(ϵ, δ) hull algorithms

April 16, 2017

Suppose we are given a point set $P \in \mathbb{R}^d$ **with diameter** ≤ 1. An ϵ-approximate convex hull is a set S such that every point in P is within Euclidean distance ϵ from some point in S. We can think of ϵ-approximate convex hull in a slightly different way.

1 Definition of (ϵ, δ) hull

Definition 1.1. Given a vector v and a point set P, we define the **directional width** as

$$\omega_v(P) = \max_{p \in P} p \cdot v$$

Definition 1.2. If p is a point we define $\omega_v(p) = p \cdot v = \omega_v(\{p\})$

It is easy to see that if $S \subseteq P$ then for all v, $\omega_v(S) \leq \omega_v(P)$.

Definition 1.3. We say S **maximizes** P in v if

$$\omega_v(P) = \omega_v(S)$$

Note that as per definition 1.1, S can be either a single vector or a set of vectors.

Definition 1.4. A **convex hull** is the minimal sized set $S \subseteq P$ such that S maximizes P in all (unit) directions v.

Definition 1.5. We say S **ϵ-maximizes** P in direction v if v *is a unit vector* and

$$|\omega_v(P) - \omega_v(S)| \leq \epsilon$$

Note that as per definition 1.1, S can be either a single vector or a set of vectors.

Definition 1.6. An ϵ**-hull** is the minimal sized set $S \subseteq P$ such that S ϵ-maximizes P in all (unit) directions v. We define OPT to be $|S|$.

Intuitively, an ϵ-approximate convex hull approximates the original point set in all directions. Coming up with a streaming algorithm that is competitive within a constant factor of OPT (the batch optimal) for this problem appears to be difficult. An interesting relaxation proposed by Lin is to have a good approximation in *most directions*. In the sections that follow, we will assume that the algorithm has access to OPT and sets $k = \text{OPT}$. In practice, we do not know OPT so we would simply set k to be the largest value our computational resources permit. We would then have an (ϵ, δ)-approximation for all point sets where $\text{OPT} \leq k$.

Definition 1.7. An (ϵ, δ)**-hull** is the minimal sized set $S \subseteq P$ such that if we pick a vector v uniformly at random from the surface of the unit sphere, $\mathcal{S}^{d-1}$, S ϵ-maximizes P in direction v with probability at least $1 - \delta$, that is,

$$\Pr(|\omega_v(P) - \omega_v(S)| > \epsilon) \leq \delta$$

Our goal is to come up with streaming algorithms for (ϵ, δ)-hulls that are competitive with OPT (the batch optimal for ϵ-hulls).

2 Core Lemmas

Definition 2.1. Define E_s^S to be the set of all vectors in $\mathbb{R}^d$ (*not just unit vectors*) that s maximizes, that is,

$$E_s^S = \{v \mid v \cdot s = \omega_v(S)\}$$

Lemma 2.1 (ϵ-Maximization Lemma)**.** *Suppose $S \subseteq P$ is an ϵ-hull of P, and fix $s \in S$. Then s ϵ-maximizes P for all unit vectors $v \in E_s^S$.*

Proof. This is because for all unit vectors v, $|\omega_v(S) - \omega_v(P)| \leq \epsilon$ and for all vectors $v \in E_s^S$, $v \cdot s = \omega_v(S)$, so for all unit vectors $v \in E_s^S$, both properties hold. $\square$

Lemma 2.2 (Covering Lemma)**.** *For all vectors $v \in \mathbb{R}^d$,*

$$v \in \bigcup_{s \in S} E_s^S$$

Proof. Given any vector v, set $s = \operatorname*{argmax}_{s' \in S} s' \cdot v$. Then $v \in E_s^S$. $\square$

Lemma 2.3 (Conic Lemma)**.** *E_s^S is a cone, that is,* *1.* $0 \in E_s^S$

2. If $v \in E_s^S$ *and* $\alpha \in \mathbb{R}^+$ *then* $\alpha v \in E_s^S$.

3. If $v, w \in E_s^S$ *then* $v + w \in E_s^S$.

Proof. We prove each item,

1. $0 \cdot s = \max_{s' \in S} 0 \cdot s' = 0$

2. $(\alpha v) \cdot s = \alpha(v \cdot s) = \alpha(\max_{s' \in S} v \cdot s') = \max_{s' \in S}(\alpha v) \cdot s'$

3. $(v + w) \cdot s = (v \cdot s) + (w \cdot s) = \max_{s' \in S} v \cdot s' + \max_{s' \in S} w \cdot s' \geq \max_{s' \in S}(v + w) \cdot s'$

□

Lemma 2.4 (Cutting Lemma)**.** *Given any 2 points* $a \neq b \in \mathbb{R}^d$, *let* $H = \{v \mid v \cdot a \geq v \cdot b\}$. *Then* H *is a closed halfspace cutting through the origin.*

Proof. Writing this in another way, $H = \{v \mid v \cdot (a - b) \geq 0\}$, which, if $a - b \neq 0$, is precisely the equation of a closed halfspace. The plane defining the boundary of the halfspace is defined by $P = \{v \mid v \cdot (a - b) = 0\}$ (that is, the set of all vectors perpendicular to $a - b$) which cuts through the origin. □

Lemma 2.5 (Bounded Maximization Lemma)**.** *Suppose* S *has at least 2 distinct points. Then if* $s \in S$, E_s^S *is contained inside a closed halfspace passing through the origin.*

Proof. Choose $s' \in S$ with $s' \neq s$. Then, let $H = \{v \mid v \cdot s \geq v \cdot s'\}$. $E_s^S \subseteq H$ but by the cutting lemma, H is a closed halfspace passing through the origin. □

3 2D-Algorithm

3.1 Algorithm

We give a deterministic algorithm that stores $O(\frac{k}{\delta})$ points and gives us an (ϵ, δ)-hull of a point set P, where k is the batch optimal for the ϵ-hull of P.

Choose $O(\frac{k}{\delta})$ equally separated unit vectors on the boundary of the unit circle. Going counter-clockwise by angle, the angle formed by any 2 consecutive vectors will be less than $\frac{2\pi\delta}{k}$. For each chosen vector v we store the point $p \in P$ s.t. $p \cdot v = \omega_v(P)$. This can be done in streaming - for a vector v, we keep an incoming point p iff $v \cdot p$ is greater than $v \cdot p'$ for the point p' we currently stored in direction v (or if we have not stored any point for direction v). Call the set of points our algorithm chooses T.

3.2 Proof

WLOG suppose that P has at least 2 distinct points (otherwise we can trivially solve the problem by storing the only point in P). Consider an optimal ϵ-hull $S \subseteq P$. WLOG suppose that S contains at least 2 points (otherwise we can simply add some point in P to S and our bounds will only change by a constant factor).

Partitioning: Pick a vector v uniformly at random on the boundary of the unit circle. By the covering lemma, $v \in E_s^S$ for some $s \in S$. Fix $s \in S$. It suffices to show the probability that $v \in E_s^S$ and T does not ϵ-maximize P in direction v is $\leq \frac{\delta}{k}$. Then, since there are k choices for s, the probability T does not ϵ-approximate P is $\leq k\frac{\delta}{k} = \delta$.

Angular setup: Fix $s \in S$. S has at least 2 distinct points, so by the cutting lemma, E_s^S is contained in a half-space. So we can rotate space such that E_s^S does not contain the positive x axis. We measure angles counter-clockwise from the positive x-axis. From the conic lemma, we know that E_s^S is the set of all vectors with angles between θ_a and θ_b with $\theta_a < \theta_b$. Out of the $O(\frac{k}{\delta})$ vectors we chose, consider the subset of vectors that are in E_s^S. Call them $v_1, v_2, ..., v_m$ and suppose they have corresponding angles $\theta_1 < \theta_2 < ... < \theta_m$. We also have that $\theta_a \leq \theta_1$ and $\theta_m \leq \theta_b$.

Lemma 3.1. *Consider some v_i. We choose a point p_i s.t. $p_i \cdot v_i = \omega_{v_i}(P)$ (that is, p_i is maximal in direction v_i). We will show that p_i ϵ-maximizes either all unit vectors with angles in the range $[\theta_a, \theta_i]$ or in $[\theta_i, \theta_b]$*

Proof. If $p_i = s$, then p_i ϵ-maximizes all unit directions in E_s^S, so we are done. Otherwise, suppose $p_i \neq s$. By the cutting lemma, there exists a closed half-space H passing through the origin, such that $p \cdot v \geq s \cdot v$ iff $v \in H$. By the ϵ-maximization lemma, for all unit vectors $v \in E_s$, $0 \leq \omega_v(P) - s \cdot v \leq \epsilon$. This implies that for all unit vectors $v \in E_s \cap H$, $0 \leq \omega_v(P) - p \cdot v \leq \omega_v(P) - s \cdot v \leq \epsilon$. In other words, p ϵ-maximizes all unit vectors $v \in E_s \cap H$.

Since E_s is itself contained in some halfspace passing through the origin, $E_s \cap H$ is either the set of vectors with angles in range $[\theta, \theta_b]$ or in range $[\theta_a, \theta]$. We note that $v_i \in E_s \cap H$ and has angle θ_i, so in either case the range contains θ_i. This proves the lemma. □

We say that v_i is *down* if p_i ϵ-maximizes P in all directions with angles in the range $[\theta_a, \theta_i]$ and *up* if p_i ϵ-maximizes P in all directions with angles in the range $[\theta_i, \theta_b]$. If v_1 is up, then p_1 ϵ-maximizes P in all directions with angles in $[\theta_1, \theta_b]$. The angle between θ_a and θ_1 is $\leq \frac{2\pi\delta}{k}$ because we chose vectors that were $\frac{2\pi\delta}{k}$ apart, so we are done. A similar argument applies if v_m is down, and if $m = 0$ (we did not choose any vectors between θ_a and θ_b. Otherwise, we consider the smallest i s.t. v_i is down but v_{i+1} is up. Then, p_i ϵ-maximizes P in all directions with angles in $[\theta_a, \theta_i]$ and p_{i+1} ϵ-maximizes P in all directions with angles in $[\theta_{i+1}, \theta_b]$. We might not ϵ-maximize P in directions with angles in the range $[\theta_i, \theta_{i+1}]$, but this angle is $\leq \frac{2\pi\delta}{k}$.

4 Generalizing to higher dimensions

4.1 Algorithm

Suppose we fix the dimension d. We give a randomized algorithm that uses n points and with probability at least $1 - p$ gives us an (ϵ, δ)-hull of a point set P, where k is the batch optimal for the ϵ-hull of P and n is defined as

$$n \in O\left(\frac{k^2}{\delta^2}\left(\log k + \log\frac{1}{\delta} + \log\frac{1}{p}\right)\right)$$

Note that the given complexity hides the dependency on d, so the actual complexity will be $O(f(d)\frac{k^2}{\delta^2}\left(\log k + \log\frac{1}{\delta} + \log\frac{1}{p}\right))$ where I think $f(d) \sim 8^d$. This section is currently a draft. The ideas are not entirely formal, but hopefully it should give a sense of what we have right now and what details we need to fill in. We think there might be a simpler proof of the algorithm, so understanding the intuition behind the algorithm might help. The bound might not be tight (or entirely correct) either.

The algorithm is simple: Choose n random vectors on the unit sphere. For each chosen vector v we store the point $p \in P$ s.t. $p \cdot v = \omega_v(P)$. As in the 2D algorithm, this can be done in streaming.

4.2 Proof Sketch

As before, WLOG suppose that P has at least 2 distinct points (otherwise we can trivially solve the problem by storing the only point in P). Consider an optimal ϵ-hull $S \subseteq P$. WLOG suppose that S contains at least 2 points (otherwise we can simply add some point in P to S and our bounds will only change by a constant factor).

Constructing small triangular cones: A triangular cone is a cone defined by the non-negative sums of 3 vectors. First, we choose c "small" triangular cones, where c is a large constant. Consider a triangle cone defined by unit vectors v_1, v_2, v_3. More precisely, we choose the cones so that $\max(|v_1 - v_2|_2, |v_2 - v_3|_2, |v_1 - v_3|_2) < 0.01$. We choose the cones so that they are disjoint, and they span $\mathbb{R}^d$, that is, every vector is in some cone. We need to prove that this is possible, but I think in any fixed dimensional space it's fairly clear that we should be able to do this. Intuitively, there should be a constructive proof, where we start out with a bunch of triangular cones that span $\mathbb{R}^d$. Then we can keep making the cones smaller (we can select a vector in the middle of each cone to split it up into 3 cones) until the cones satisfy the required condition.

Bounding size of E_s cones: Next, we consider the intersection of each cone E_s with each of the c triangular cones. This means we get a total of ck cones. The cones are small, in the sense they are contained in the "small" triangular cones we chose. Consider the intersection of any of these cones with the unit sphere. If the measure of the intersection is small, that is, if it is smaller than $\frac{\delta}{2ck}$ then we can ignore the cone. This is because even if we don't ϵ-maximize directions in these cones, we have at most ck such cones, so we miss at most $\frac{\delta}{2}$ of all directions. In other words, WLOG the cones are not "too small".

Cutting cones: Consider a cone $C \cap E_s^S$. Intuitively, we are going to show that each time we choose a vector v, we cut away part of $C \cap E_s^S$ (for the part we cut off, we have an ϵ-maximization). Consider each unit vector v we choose in cone $C \cap E_s^S$, and corresponding point p that maximizes P in v. If $p = s$, then p ϵ-maximizes the whole cone E_s^S so we are done. Otherwise, by the cutting lemma, there exists a closed half-space H passing through the origin, such that $p \cdot v' \geq s \cdot v'$ iff $v' \in H$. As in the 2D proof, p then ϵ-maximizes all vectors in $E_s^S \cap H$ and therefore in $C \cap E_s^S \cap H$. So the region of $C \cap E_s^S$ that we might not have ϵ-maximized is $C \cap E_s^S \cap H^c$. $C \cap E_s^S \cap H^c \subseteq C \cap E_s^S \cap (H^c \cup \{0\})$. $(H^c \cup \{0\})$ is a cone, and the intersection of cones is a cone, so $C \cap E_s^S \cap (H^c \cup \{0\})$ is also a cone. Note that $v \in H$ and $v \neq 0$, so $v \notin C \cap E_s^S \cap (H^c \cup \{0\})$.

Projection: After choosing many vectors $v \in C \cap E_s^S$, we end up with a cone $C' \subseteq C \cap E_s^S$ where C' does not contain any of the selected vectors v. We want to show that the area of $C' \cap \mathbb{S}^2$ is less than $\frac{\delta}{ck}$ fraction of the unit sphere's area. The main trick will be to project this into a 2-dimensional problem. Suppose that (triangular) cone C was defined by vectors v_1, v_2, v_3. Consider the plane P containing the endpoints of v_1, v_2, v_3. We project $C \cap E_s^S \cap \mathbb{S}^2$ onto P, and similarly project all unit vectors we chose that are in $C \cap E_s^S \cap \mathbb{S}^2$ onto P, and project $C' \cap \mathbb{S}^2$ onto P.

Reduced problem: This gives us a simpler problem. Suppose we have convex polygon C with area A, where we choose points in C from a nearly uniform distribution π. More formally, the ratio between the max and min of the PDF of π is at most 2. Call a convex polygon *respectful* if it does not contain any selected points. How many points do we need to choose so that with high probability, all respectful convex polygons contained in C must have area $\leq \frac{\delta}{k}$? The reason this reduction holds is because we chose cone C to be small, so $C \cap \mathbb{S}^2$ is quite flat. In particular, if we rotate space so that plane P is perpendicular to the z axis, the norm of the gradient of $C \cap \mathbb{S}^2$ is bounded by 2. So when a set of measure v in $C \cap E_s^S \cap \mathbb{S}^2$ is projected to P it has measure between $\frac{v}{2}$ and v.

Lemma 4.1. *Given a convex polygon C, there exists 3 points in C such that the triangle formed between them has area at least $\frac{1}{4}$ the area of C.*

Lemma 4.2. *The pdf of π must be between $\frac{1}{2A}$ and $\frac{2}{A}$ everywhere.*

Uniform Bound Outline Let T be the set of triangles contained in polygon C with area $\geq \frac{\delta}{4ck}$. Given $t \in T$, let $P(t)$ denote the probability that a point selected from π lies in t. Since π is almost uniform, $P(t) \geq \frac{\delta}{8Ack}$. We will use uniform bounds and VC dimension, to show that if we select enough points in C then with high probability *every* triangle in T will contain a point. This is sufficient to solve the problem: consider arbitrary convex polygon G contained in C with area $\geq \frac{\delta}{ck}$. By lemma 4.1, it contains a triangle of area $\geq \frac{\delta}{4ck}$. But all such triangles contain a point, so G contains a point. Therefore every convex polygon in C that does not contain a point must have area $< \frac{\delta}{ck}$.

Sampling Process Select n points in C from distribution π. For $t \in T$, let $C_n(t)$ be the number of selected points that lie in t, and let $P_n(t)$ be $\frac{C_n(t)}{n}$. Intuitively, we are estimating the cumulative density of π in a region by sampling points and computing the proportion of points that fall in the region. Suppose that for all $t \in T$, $|P_n(t) - P(t)| < \frac{\delta}{8Ack}$ (this means that P_n approximates P well). Then, for all t, since $P(t) \geq \frac{\delta}{8Ack}$, $P_n(t) > 0$. This means that every triangle $t \in T$ contains a point.

VC Dimension We want to show that with probability at least $1-\frac{p}{2ck}$, $\sup_{t\in T} |P_n(t) - P(t)| < \frac{\delta}{8Ack}$. The VC dimension of T is 7. Letting $\epsilon = \frac{\delta}{8Ack}$, we apply a VC-dimension based uniform bound to get that,

$$P(\sup_{t\in T} |P_n(t) - P(t)| > \epsilon) \leq 8(n+1)^7 e^{-n\epsilon^2/32}$$

Setting $n \in O(\frac{A^2k^2}{\delta^2}(\log \frac{Ak}{\delta} + \log \frac{k}{p}))$ works, since $k > 2$ and $\delta < 1$.

Finishing up: Consider a cone $C \cap E_s^S$. Suppose that the intersection $C \cap E_s^S \cap \mathbb{S}^2$ has area A. The projection has area $\leq A$, so it suffices to choose $n \in O(\frac{A^2k^2}{\delta^2}(\log \frac{Ak}{\delta} + \log \frac{k}{p}))$ unit vectors in $C \cap E_s^S \cap \mathbb{S}^2$. When we choose a unit vector at random, it lies in $C \cap E_s^S \cap \mathbb{S}^2$ with probability $c = \frac{A}{4\pi}$. We want to choose m points, so that with probability at least $1 - \frac{p}{2ck}$ we have n points inside the cone. From a standard Chernoff bound argument, $m \in O(\frac{n}{c}(\log n + \log \frac{k}{p}))$ works. Simplifying, this becomes,

$$\begin{aligned} m &= O\left(\left[\left(\frac{A^2k^2}{\delta^2}\left(\log\frac{Ak}{\delta} + \log\frac{k}{p}\right)\right)\left(\frac{4\pi}{A}\right)\right]\left[\log n + \log\frac{k}{p}\right]\right) \\ &= O\left(\left[\left(\frac{Ak^2}{\delta^2}\left(\log\frac{Ak}{\delta} + \log\frac{k}{p}\right)\right)\right]\left[\log n + \log\frac{k}{p}\right]\right) \end{aligned}$$

We note that $A \leq 4\pi$. Furthermore, $\log n$ actually simplifies to

$$O(\log\frac{k}{\delta} + \log(\log\frac{k}{\delta} + \log\frac{k}{p})) \subseteq O(\log\frac{k}{\delta} + \log\frac{k}{p})$$

So in the worst case m reduces to

$$m \in O\left(\frac{k^2}{\delta^2}\left(\log k + \log\frac{1}{\delta} + \log\frac{1}{p}\right)\right)$$

We can apply union bounds over the cones to get a $\leq p$ chance of failure.

4.3 Dealing with higher dimensions

A few parts of the proof need to be verified and augmented to deal with higher dimensions.

1. The section where we construct small triangular cones. In d dimensional space, the cones will be defined by d vectors such that the pairwise distance between any 2 vectors is at most 0.1. We need to prove that this can be done, and analyze how the number of cones needed scales with d.

2. We need to verify that the projection to a $d-1$ dimensional problem works.

3. We need to modify lemma 4.1 to something like: every d dimensional polyhedra C has a subset of d points S, such that the ratio of the d-dimensional volume of C to the d-dimensional volume of S is at most 2^{d-1}.

4. We need to verify and bound the VC dimension of a d dimensional shape containing d points.