COMPSCI 634: Geometric Algorithms March 31, 2016

Lecture 19

Lecturer: Pankaj Agarwal *Scribe: Allen Xiao*

1 Overview

In the last lecture, we showed how ϵ-net theory produced better range searching data structures. In this lecture, we continue our investigation of geometric summaries via sampling. We start by demonstrating how random sampling fails to preserve certain geometric properties, then define a class of properties where where *coresets* succeed.

2 Motivating example

Suppose we have a set S of n points in $\mathbb{R}^d$, and an *optimization problem* $\mathbb{P}(S)$ over S. For example, $\mathbb{P}(S)$ could optimize for

- the farthest pair (diameter),
- the smallest enclosing ball, box, ellipsoid, etc.,
- the minimum spanning tree,
- or a clustering of S.

Many of these problems are challenging even for moderate dimension ($d > 2$). Instead of exact solutions, we look towards approximation. Let $\mu(S)$ be the cost of the optimal solution to $\mathbb{P}(S)$; our goal is to find a solution which ϵ-approximates $\mu(S)$. With the ideas of the past few lectures, a natural strategy for approximation is to find a small, representative sample of $R \subseteq S$, and solve $\mathbb{P}(R)$. Formally, we introduce the notion of a *coreset*.

Definition 1. *A subset $R \subseteq S$ is an **ϵ-coreset** of S for measure $\mu(\cdot)$ if*

$$|\mu(R) - \mu(S)| \leq \epsilon\mu(S).$$

We applied this successfully for ϵ-nets/samples, but it is unclear whether there exists a small coreset $R \subseteq S$, and even if R exists, that we can compute it quickly. In fact, the random sampling we used for ϵ-nets/samples seems to perform poorly on the $\mathbb{P}(S)$ examples above. Intuitively, random sampling preserves *statistical* properties, but not necessarily *metric* properties.

Example 1. Let $\mathbb{P}(S)$ be the diameter. Very few ($O(1)$) points are needed to define the diameter of S. If the random sample discards them, ϵ-approximation may be hopeless.

Even the "capture large ranges" notion from ϵ-samples is insufficient for some problems. For example, the smallest enclosing ball of an ϵ-sample need not contain all the original points (so such a ball is infeasible).

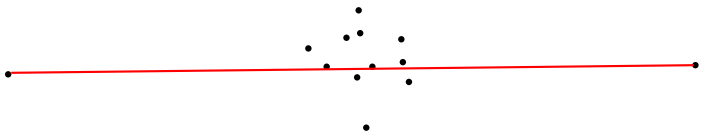

Figure 1: Only one pair of points support the diameter.

3 Coresets for faithful measures

We introduce a surrogate to random sampling which apply under specific conditions on $\mu(S)$, but still generalize to a wide variety of geometric problems. Many problems, including most of the examples listed earlier, have the following property.

Definition 2. *A measure $\mu(\cdot)$ is **faithful** if, for all S, $\mu(S) = \mu(\text{conv}(S))$.*

Intuitively, the convex hull contains enough structure to evaluate the measure. Returning to the example, we see that diameter is indeed faithful.

Our roadmap for the next few sections is the following: we will show that there is a small, quickly computable subset for approximating $\text{conv}(S)$, and that this subset also serves as a coreset for *any faithful measure*. To formalize the notion of approximating the convex hull, we first introduce directional width.

Definition 3. *Let a **direction** be any unit vector in $\mathbb{R}^d$ (i.e. $u \in \mathbb{S}^{d-1}$). Given direction u, the **directional width** of S in direction u is*

$$w(S, u) = \max_{p \in S} \langle p, u \rangle - \min_{q \in S} \langle q, u \rangle.$$

Visually, consider two hyperplanes orthogonal to $u \in \mathbb{S}^{d-1}$ squeezing S from opposite sides, stopping on the first points they touch. The directional width is the distance between them when they stop.

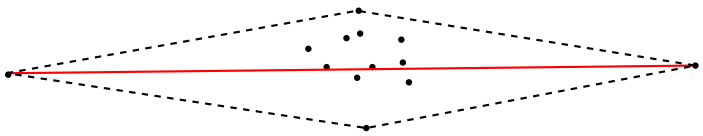

Figure 2: $\text{diam}(\text{conv}(S))$ is same as $\text{diam}(S)$.

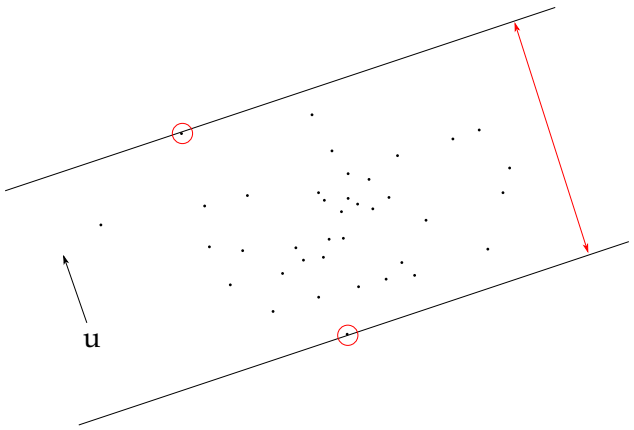

Figure 3: Directional width.

Remark 1. *The **width** of S is*

$$\min_{u \in \mathbb{S}^{d-1}} w(S, u).$$

Formally, we will call the subset we find an ϵ-kernel.

Definition 4. *Let $R \subseteq S$. R is an **ϵ-kernel** of S if, for all directions u,*

$$w(S, u) \geq w(R, u) \geq (1 - \epsilon) w(S, u).$$

Intuitively, such an R must have $\text{conv}(R) \sim \text{conv}(S)$ in all directions. Once we have an ϵ-kernel, we can state the first main theorem.

Theorem 1. *If R is an ϵ-kernel of S, then R is a $(c\epsilon)$-coreset for any faithful measure,*

$$|\mu(R) - \mu(S)| \leq c\epsilon\mu(S),$$

where c is a constant which may depend on d.

The second main theorem answers our earlier questions of computation time and size.

Theorem 2. *Given set $S \subset \mathbb{R}^d$ of n points and $\epsilon > 0$, an ϵ-kernel of S of size $O(1/\epsilon^{(d-1)/2})$ can be computed in time $O(n + 1/\epsilon^{d-1})$.*

Importantly, the size of an ϵ-kernel depends only on ϵ and d, and not n. This size bound is actually tight in the worst case. Consider an S of points arranged evenly around a circle in $\mathbb{R}^2$. In such an arrangement, $1/\sqrt{\epsilon}$ samples are necessary for the ϵ error in directional width.

Lastly, the size depends exponentially on d, and in practice this performs well up to $d = 8 \sim 10$. Applying this theorem to obtain a coreset is straightforward, as we can see in this example for diameter.

Example 2.

Let $R \subseteq S$ be an ϵ-kernel for S, and suppose we approximate diameter by naively computing the farthest pair in R (test all pairs).

$$\operatorname*{argmax}_{p,q \in R} \|p - q\|$$

By Theorem 2, computing R takes $O(n + 1/\epsilon^{d-1})$ time. Due to the small size of R, brute-force search for the farthest pair only takes $O(1/\epsilon^{d-1})$ time. The total time is only $O(n + 1/\epsilon^{d-1})$.

Claim 1. *The ϵ-kernel R is an ϵ-coreset for diameter, i.e.*

$$\operatorname{diam}(R) \geq (1-\epsilon)\operatorname{diam}(S).$$

Proof. Let p^*, q^* be the farthest pair of S, realizing the diameter. Let direction $u^* = \frac{q^*-p^*}{\|q^*-p^*\|}$, and consider the strip containing S orthogonal to u^*. Since R is an ϵ-kernel, there exists some pair $a^*, b^* \in R$ which ϵ-approximates the directional width in u^*. By definition of directional width,

$$\operatorname{diam}(R) \geq \|a-b\| \geq w(R,u^*) \geq (1-\epsilon)w(S,u^*) = (1-\epsilon)\operatorname{diam}(S),$$

and R is an ϵ-coreset of S for diameter. □

In the next section, we will describe a slower algorithm for constructing an ϵ-kernel. Afterwards, we will sketch the construction achieving Theorem 2. Both will be described without thorough analysis. A detailed treatment (with analysis) can be found in Chapter 23 of Har-Peled's book [HP11].

4 Constructing an ϵ-kernel

We begin with a weaker construction, which highlights some of the key ideas. Without loss of generality, assume that the smallest enclosing hypercube of S has unit side length (this can be achieved by performing an affine transformation on S).

1. Draw a grid of cell-size $\epsilon/10$ over the hypercube.
2. In each column of the d-th dimension, choose an arbitrary point from
   - the topmost nonempty cell and
   - the lowest nonempty cell.

   to form R. The size of R is therefore $O(1/\epsilon^{d-1})$.

Claim 2. *The subset R is an ϵ-kernel.*

Informally, the argument is as follows. For any direction u, the directional width is supported by two points p^*, q^* from cells which we've sampled points for R. The diameter of each cell is $O(\epsilon)$, which means that there are points of R within $O(\epsilon)$ of p^*, q^*, therefore the directional width of R is within some $O(\epsilon)$ of S.

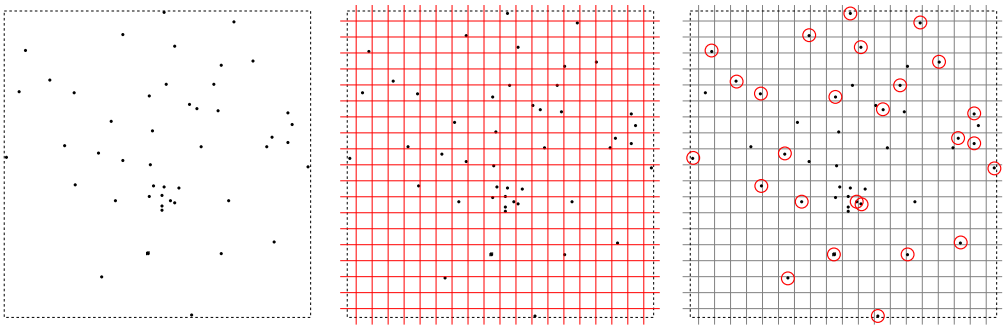

Figure 4: Section 4's construction of an ϵ-kernel.

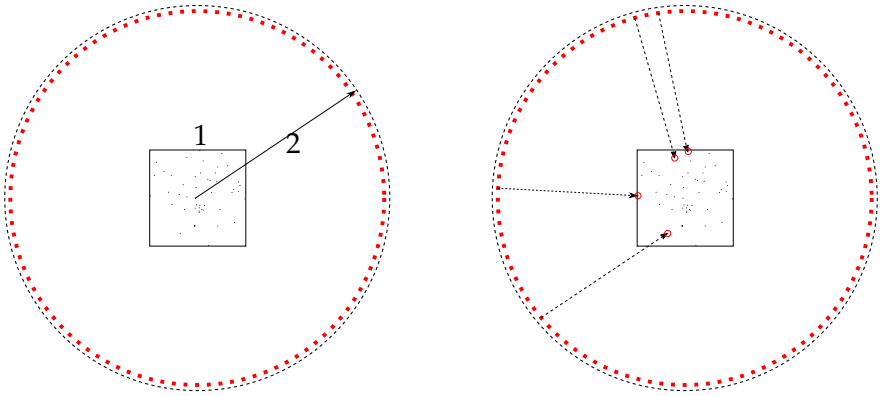

Figure 5: Some steps of the Theorem 2 construction.

4.1 Kernel construction for Theorem 2

Again, consider the unit hypercube acting as the minimum enclosing hypercube of S.

1. Assume that this hypercube is centered at the origin, and draw an origin-centered ball of radius 2 about it.
2. Let C be a set of $O(1/\epsilon^{(d-1)/2})$ points uniformly spaced about the surface of the ball.
3. Add the nearest neighbor (from S) of each point of C to R.

$$R = \{\text{NN}(q, S) \mid q \in C\}$$

In practice, the ϵ-approximate nearest neighbor is chosen instead of the exact.

The central idea is similar to the previous algorithm. For a fixed direction u, let the maximizing point of $\max_{p \in S} \langle u, p \rangle$ be $p^* \in S$ (i.e. p^* realizes one side of the directional width in u). Let $q = \text{NN}(p^*, C)$, then the point $\text{NN}(q, S)$ we selected for the kernel must lie sufficiently close to p^* such that the projections p^* and $\text{NN}(q, S)$ onto u differ by at most ϵ.

5 Summary

We observed that certain geometric measures were approximated poorly using random samples, and formalized this notion of approximation-by-subset using coresets. Then, we showed that ϵ-kernels would function as coresets for faithful geometric measures. Finally, we described a construction for a small ϵ-kernel.

Good coreset constructions exist for other measures, such as sets of functions and clustering. For examples, see the survey by Agarwal et al. [AHPV05]. The original paper [AHPV04] is by the same authors. The faster ϵ-kernel construction is due independently to Chan [Cha04] and Yu et al. [YAPV08].

References

[AHPV04] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Approximating extent measures of points. *Journal of the ACM (JACM)*, 51(4):606–635, 2004.

[AHPV05] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. In Jacob E. Goodman, Janos Pach, and Emo Welzl, editors, *Combinatorial and Computational Geometry*, volume 52, pages 1–30. Cambridge University Press, New York, 2005.

[Cha04] Timothy M. Chan. Faster core-set constructions and data stream algorithms in fixed dimensions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, pages 152–159. ACM, 2004.

[HP11] Sariel Har-Peled. *Geometric approximation algorithms*, volume 173. American Mathematical Society, Providence, 2011.

[YAPV08] Hai Yu, Pankaj K. Agarwal, Raghunath Poreddy, and Kasturi R. Varadarajan. Practical methods for shape fitting and kinetic data structures using coresets. *Algorithmica*, 52(3):378–402, 2008.