CSE 522: Sublinear (and Streaming) Algorithms **Spring 2014**

Lecture 13: Coresets and Geometric Streams

May 12, 2014

Lecturer: Paul Beame *Scribe: Paul Beame*

1 Approximate Ranking

Before we move on, we briefly mention the last of the statistical problems we will consider for the data stream model. The problem in this case is to approximately produce the rank of elements in the input sequence. The relative rank of j with respect to a data stream with non-negative input frequencies is $\sum_{i \le j} f_i / ||f||_1$. There are two versions of the question that have similar solutions: Given j, estimate the relative rank of j within ε, or given a relative rank $R \in [0, 1]$, produce an element from the input stream whose relative rank is within ε of R.

In the first problem on problem set 2 you will analyze a general method for approximate ranks based on the heavy hitters algorithms we have discussed. If only insertions are allowed, there is a more space-efficient algorithm due to Greenwald and Khanna that uses somewhat related ideas and would be a better choice in practice.

2 Geometric Streams and Coresets

We now consider streams where the input domain has more structure than simply $[M]$. In particular, we consider streams of points from $\mathbb{R}^d$ for small d. In this context we will have to modify our notion of space, since even a single value in $\mathbb{R}$ would require an unbounded number of bits. We define Space to be the number of input points we keep (plus whatever bits we need).

A notion we will find particularly useful is that of a *coreset* for a point set relative to a cost function. Coresets (a) yield enough information to approximately determine the optimum solution on the set of points, but (b) "compose" nicely.

One simple problem that we will consider is the *Minimum Enclosing Ball (MEB)* Problem, which requires one to find the minimum radius Euclidean ball that contains the input point set: Given a point set $P \subseteq \mathbb{R}^d$,

$$MEB(P) = \inf_{x \in \mathbb{R}^d} \max_{y \in P} ||y - x||_2.$$

A *cost function* C is a map from the input space $\mathbb{R}^d$ to the non-negative reals, parameterized by sets

of points $P \subset \mathbb{R}^d$, $C_P : \mathbb{R}^d \to \mathbb{R}_+$. We will sometimes allow the cost function to be parameterized by weighted point sets that have a non-negative weights on each point.

For example, the MEB Problem has a natural associated cost function over potential centers $x \in \mathbb{R}^d$ given by

$$C_P(x) = \max_{y \in P} ||y - x||_2.$$

The natural functions that coresets will allow us to approximate is $\inf_x C_P(x)$ as well as finding an x for which that minimum is approximated.

We note that for the MEB problem the function C_P is *monotone* in P, in that for $P \subseteq Q$, $C_P(x) \leq C_Q(x)$ for all $x \in \mathbb{R}^d$.

Definition 2.1. *For any real number $\alpha \geq 1$ and cost function C parameterized by (weighted) point sets in $\mathbb{R}^d$ is an* α-coreset *for $P \subset \mathbb{R}^d$ with respect to C if $S \subseteq P$ and for all $T \subseteq \mathbb{R}^d$ and all $x \in \mathbb{R}$,*

$$C_{S \cup T}(x) \leq C_{P \cup T}(x) \leq \alpha\, C_{S \cup T}(x).$$

The key condition for a coreset is that it not only provides a summary of set P itself with respect to approximation cost, it also provides summary of P with respect to all future extensions of P. It is immediate that if C_P is monotone in P, then first of the two inequalities is automatically satisfied and it is only the upper bound of $C_{P \cup T}$ by $\alpha\, C_{C \cup T}$ that needs to be argued.

We will show that the cost function C^{MEB} for MEB has $(1+\varepsilon)$-coresets of size $O(1/\varepsilon^{(d-1)/2})$ for every point set P. We will use this together with a generic streaming algorithm based on coresets to derive a small space data streaming algorithm for the $(1+\varepsilon)$-approximate MEB Problem.

Before we give this construction we give a simpler example of coresets.

Coresets for the Median We consider sets of points P in $\mathbb{R}$ and cost function $C_P(x) = \max(\#\{a \in P \mid a < x\}, \#\{a \in P \mid x < a\})$ which is minimized if x is the median of P. A set $S \subseteq P$ that yields a $(1+\varepsilon)$-approximation to $C_P(x)$ produces a minimum that is an element whose relative rank is $1/2 \pm \varepsilon/2$ where $n = |P|$. If the elements of P are $a_1 \leq a_2 \leq \cdots \leq a_n$ then the coreset S for P will consist of $1/\varepsilon$ elements $a_{\varepsilon n}, a_{2\varepsilon n}, \ldots, a_n$ each with weight εn.

Merge Property of Coresets If S is an α-coreset for P and S' is a β-coreset for Q then $S \cup S'$ is an $\alpha\beta$-coreset for $P \cup Q$.

Proof. We have

$$C_{S \cup S' \cup T}(x) \leq C_{P \cup S' \cup T}(x) \leq C_{P \cup Q \cup T}(x) \leq \alpha\, C_{S \cup Q \cup T}(x) \leq \alpha\beta\, C_{S \cup S' \cup T}(x)$$

where the first and third inequalities follow from the α-coreset property of S for point set P using sets $T' = S' \cup T$ and $T'' = Q \cup T$, and the second and fouth inequalities from the β-coreset property of S' for Q using sets $T''' = P \cup T$ and $T'''' = S \cup T$. □

Reduce Property of Coresets If S is an α-coreset for P and S' is a β-coreset for S then S' is an $\alpha\beta$-coreset for P.

Proof. We have

$$C_{S' \cup T}(x) \leq C_{S \cup T}(x) \leq C_{P \cup T}(x) \leq \alpha\, C_{S \cup T}(x) \leq \alpha\beta\, C_{S' \cup T}(x)$$

□

Definition 2.2. *We say that a cost function C has the* disjoint union property *iff whenever S is an α-coreset for P, S' is an α-coreset for Q, and $P \cap Q = \emptyset$ then $S \cup S'$ is an α-coreset for $P \cup Q$.*

The MEB cost function clearly satisfies the disjoint union property but not all cost functions satisfy this property.

The coreset construction for MEB will be based on a relatively small collection of directions in $\mathbb{R}^d$ that approximate every possible direction in $\mathbb{R}^d$.

Definition 2.3. *For an angle $\theta > 0$, we call a collection $V = \{v_1, \ldots, v_t\} \subset \mathbb{R}^d$ of vectors in a* θ-net *iff every $u \in \mathbb{R}^d$ there is a $v_i \in V$ such that the angle between u and v_i is at most θ.*

Theorem 2.4. *For every $\theta > 0$ there is θ-net consisting of $O(1/\theta^{d-1})$ vectors.*

Proof Sketch. The set of vectors within angle $\theta < 1$ of a given vector on the unit sphere covers a patch on the surface of the unit ball of area $\Omega(\theta^{d-1})$ since $\sin\theta$ is $\Omega(\theta)$ and the surface is of dimension $d-1$. We choose the θ-net to be a maximal collection of vectors on the unit sphere no two of which have angle less than θ. Observe that the sets of vectors within angle less than $\theta/2$ from each of these given vectors form disjoint regions on the surface of the sphere and each have area $\Omega(\theta^{d-1})$. Since the unit sphere has constant surface area, there are at most $O(1/\theta^{d-1})$ such vectors in the θ-net. □

From this we obtain coresets of size independent of the size of the point sets from which they are derived.

Theorem 2.5. *In $d \geq 2$ dimensions the MEB cost function has $(1+\varepsilon)$-coresets of size $O(1/\varepsilon^{(d-1)/2})$.*

Proof. Let $V = \{v_1, \ldots, v_t\}$ be a θ-net for $\mathbb{R}^d$ with $\theta = \sqrt{\varepsilon}$. Let $P \subset \mathbb{R}^d$. The coreset S for P for the MEB cost function will be the set of at most $2t$ points

$$\arg\max_{y \in P} \langle v_i, y \rangle \text{ and } \arg\min_{y \in P} \langle v_i, y \rangle$$

over all vectors $v_i \in V$. By construction its size is $O(1/\theta^{d-1}) = O(1/\varepsilon^{(d-1)/2})$ as claimed. The set S determines a set of t pairs of parallel hyperplanes that sandwich the points in P in each of the t directions given by V.

It remains to prove that S is a $(1+\varepsilon)$-coreset for P. Let $T \subseteq \mathbb{R}^d$ and $x \in \mathbb{R}^d$.

Since the cost function for MEB is monotone $C_{S \cup T}(x) \leq C_{P \cup T}(x)$.

It remains to show that $C_{P \cup T}(x) \leq (1+\varepsilon) C_{S \cup T}(x)$. Let

$$z = \arg\max_{y \in P \cup T} ||y - x||_2$$

which implies that $C_{P \cup T}(x) = ||z - x||_2$.

If $z \in T$ then $C_{P \cup T}(x) = ||z - x||_2 \leq C_{S \cup T}(x)$.

If $z \in P$, consider the vector xz and let $\pm v_i$ be a vector in V that makes angle at most θ with xz. Let z' be the projection of z on the ray R through x in direction $\pm v_i$. Observe that the triangle formed by x, z, and z' is a right-angled triangle with hypotenuse xz and the angle between xz and xz' is at most θ.

Since z is a candidate for $\arg\max_{y \in P} \langle v_i, y \rangle$ and $\arg\min_{y \in P} \langle v_i, y \rangle = \arg\max_{y \in P} \langle -v_i, y \rangle$. there is an element of S whose projection on R is at least as far along R as z' is and hence has distance further than $||z' - x||_2$. Therefore

$$\begin{aligned} C_{S \cup T}(x) &\geq C_S(x) \geq ||z' - x||_2 \\ &= ||z - x||_2 \cos\phi \\ &\geq ||z - x||_2 \cos\theta \\ &\geq C_{P \cup T}(x)\theta \end{aligned}$$

where $\phi \leq \theta$ is the angle between xz and xz'. Now the Taylor series for $\cos\theta = 1 - \theta^2/2! + \theta^4/4! - \theta^6/6! + \cdots \geq 1 - \theta^2/2$ for $\theta \leq 1$. Therefore

$$\begin{aligned} C_{P \cup T}(x) &\leq \frac{1}{\cos\theta} C_{S \cup T}(x) \\ &\leq \frac{1}{1 - \theta^2/2} C_{S \cup T}(x) \\ &\leq (1 + \theta^2/2) C_{S \cup T}(x) \\ &< (1+\varepsilon) C_{S \cup T}(x) \end{aligned}$$

as required. □

Next time we will show a generic method for converting coreset constructions like these to small space streaming algorithms for approximating $\inf_x C_P(x)$.