# Lecture 12: Coresets

Lecturer: *Przemysław Uznański*                                          Scribe: -

## 1   Coresets [Che09]

Setup: given set of $P \subseteq \mathbb{R}^d$, compute $C_P() : \mathbb{R}^d \to \mathbb{R}$. For example:

- $MEB$, a minimal enclosing ball: $C_P(o)$ is a radius of minimal enclosing ball for set $P$ centered at $o$.

$|P| = n$ is large, so storing it explicitly is out of the question. Observe that for fixed $o$, a single $p \in P$ is enough, that is $\forall_o \exists_{p \in P} C_P(o) = C_{\{p\}}(o)$. Can we generalize this observation so it works for all $o \in \mathbb{R}^2$ at once?

**Definition 1.** *We say that $S \subseteq P$ is a c-coreset for $P$ if for any o and any $T \subseteq \mathbb{R}^d$ there is:*

$$C_{S \cup T}(o) \leq C_{P \cup T}(o) \leq c \cdot C_{S \cup T}(o)$$

Note: this is a stronger definition than just requiring that $f_S(o)$ is a $c$-approximation to $C_P(o)$ (take $T = \emptyset$).

Observe that in case of $MEB$ we have for $A \subseteq B$: $C_A(o) \leq C_B(o)$, so the first inequality is "for free".

**Lemma 2** (Merge property). *If $S$ is a c-coreset for $P$, and $S'$ is a c-coreset for $P'$, then $S \cup S'$ is a $c^2$-coreset for $P \cup P'$.*

**Lemma 3** (Reduce property). *If $S$ is a c-coreset for $P$ and $P$ is a c-coreset for $Q$, then $S$ is a $c^2$-coreset for $Q$.*

We sometimes want a stronger property (which for example MEB satisfies)

**Property 4** (Disjoint merge). *If $S$ is a c-coreset for $P$ and $S'$ is a c-coreset for $P'$ and $P \cap P' = \emptyset$, then $S \cup S'$ is a c-coreset for $P \cup P'$.*

Exercise: proof for MEB.

**Theorem 5.** *Assume that a problem is supported by a $(1 + \alpha)$-coreset of size $f(\alpha)$, computable in linear space, with disjoint merhe property. Then there is a streaming algorithm with $1 + \varepsilon$ guarantee, with space $\mathcal{O}(f(\varepsilon/\log n) \log n)$.*

*Proof.* Sketch of a proof: put stream of $n$ elements into binary tree. Each node stores coreset for the range below it.

For two sibling nodes $N_1, N_2$ covering sets $A_1, A_2$, at level $i$, and parent node $N$ at level $i + 1$, there is:

- $N_1$ is $(1 + \alpha)^i$-coreset for $A_1$ (and the same for $N_2$, $A_2$)

- $N_1 \cup N_2$ is $(1+\alpha)^i$-coreset for $A_1 \cup A_2$

- $N$ is constructed as $(1+\alpha)$-coreset for $N_1 \cup N_2$

- thuse $N$ is $(1+\alpha)^{i+1}$-coreset for $A_1 \cup A_2$

As a end-result, we have in the root $(1+\alpha)^{\log n}$-coreset for whole input. Selecting $\alpha = \mathcal{O}(\varepsilon/\log n)$ is enough. $\qquad\square$

## 1.1 Coreset for MEB

Construction goes as follow. Choose dense set of directions $\{v_i\}_{i=1}^m$, such that for any other direction $u$, there is always some $v_i$ such that $\mathrm{angle}(u, v_i) \leq \alpha$: this is $\sim \alpha$-net on unit-ball (up to trigonometry). We can choose such set of $m = (1/\alpha)^{\mathcal{O}(d)}$ directions.

**Claim 6.** *For any direction $v_i$, pick $p_i \in P$ that is extremal in that direction. Set $S = \{p_i\}_{i=1}^m$ is a $(1 + \mathcal{O}(\alpha^2))$-coreset for $P$.*

*Proof.* Pick arbitrary $T$ ($T = \emptyset$ w.l.o.g. in this proof), and $P$ and constructed set $S$. Fix $o$. Pick furthest point $x \in P$, and close direction $v_i$, and maximal in this direction point $p_i \in S$. The angle $x$-$o$-$p_i$ is small (at most $\alpha$), so the stretch is upperbounded by $\frac{1}{\cos\alpha} = 1 + \mathcal{O}(\alpha^2)$. $\qquad\square$

## 1.2 Coreset for median

Approximate median: given sequence of numbers $A$ of number $a_1, \ldots, a_n$, return $a$ such that $(1/2 \pm \varepsilon)n$ elements in $A$ are smaller/larger than $a$.

Alternative formulation: find $a$ that minimizes $C_A(a) = \max(|\{i : a_i \geq a\}|, |\{i : a_i \leq a\}|)$.

Coreset of size $1/\varepsilon$: pick every $\varepsilon n$ element from sorted $A$. Easy to see that

$$C_{A \cup T}(a) \leq C_{A \cup S}(a) \leq (1+\varepsilon)C_{A \cup T}(a)$$

Plugging into the theorem, we obtain streaming median computation in space $\mathcal{O}(\log^2 n/\varepsilon)$. In fact this works for any quantile computation (but the error is additive). Improvement: pre-filter and keep only $1/\varepsilon^2$ elements (randomly), so space becomes $\mathcal{O}(\log^2(1/\varepsilon)/\varepsilon)$.

# 2 Graph algorithms

## 2.1 Certificates

Graph-theoretic approach, for decision problems.

**Definition 7.** *For property $\mathcal{P}$, and a graph $G$, we say that $G'$ is a strong certificate for $G$ if: for any $H$, $G \cup H$ is in $\mathcal{P}$ iff $G' \cup H$ is in $\mathcal{P}$.*

Examples:

- connectivity: any spanning forest

- non-bipartiteness: any spanning forest + single odd-cycle inducing edge

- edge/vertex connectivity

## 2.2 Spanners [**FWY19**]

Subgraph approximately preserving distances:

**Definition 8.** $H$, an edge subgraph of $G$, is a $t$-spanner of $G$, if for any $u, v$ there is

$$d_H(u, v) \leq t \cdot d_G(u, v)$$

**Theorem 9.** Any unweighted $G$ contains $t$-spanner with $\mathcal{O}(n^{1+2/(t-1)})$ edges, and it can be computed in one pass.

*Proof.* Maintain subgraph. Process edges one-by-one. If an edge causes a cycle of length $\leq t + 1$, do not insert it.

1. This constructs a $t$-spanner. (Nice easy exercise)

   *Proof.* Let's denote a spanner by $H$ as previously.

   - (by contradiction) Assume that $H$ maintained in that way is not a $t$-spanner. So, there exist a pair of vertices $u, v$ (counterexample) s.t. $d_H(u, v) > td_G(u, v)$. Let's choose some shortest path between $u, v$ in G, $P_G$ and in $H$, $P_H$. The proof will have three stages.
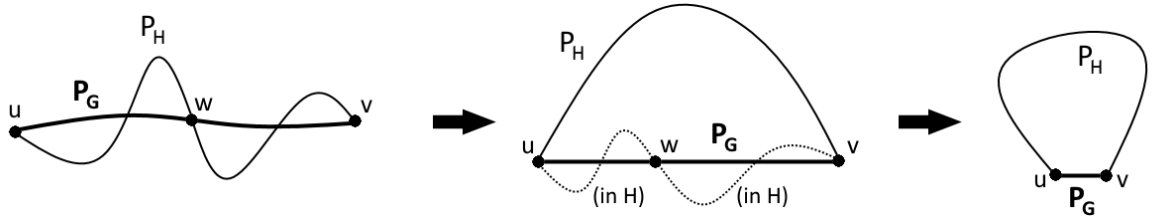


   - Now we argument that there also exists a counterexample s.t. $P_G$ and $P_H$ don't cross - only common vertices between them are $u, v$. This follows, because if there is a vertex $w$ s.t. $w \neq u, w \neq v, w \in P_G, w \in P_H$, then we can substitute one of $u, v$ (wlog, $u$) with $w$ in the counterexample. That's because $P_G, P_H$ are shortest paths between $u, v$ in $G, H$, so the equalities below are true (if they wouldn't be, there would exist some shorter path):

$$d_H(u, w) + d_H(w, v) = d_H(u, v) > td_G(u, v) = td_G(u, w) + td_G(w, v)$$

   so either $d_H(u, w) > td_G(u, w)$ or $d_H(w, v) > td_G(w, v)$ (in other case contradiction by summing sides of opposite ($\leq$) inequalities). By substituting $u \leftarrow w$ as many times as needed we have $P_G \cap P_H = \{u, v\}$.

   - We have that $P_G, P_H$ only have $u, v$ common and now we argument that there also exists a counterexample s.t. $|P_G| = 1$. Similarly as previously, if there exists some $w$ on $P_G$ between $u, v$, we can substitute one of $u, v$ (wlog, $u$) with it in the counterexample. This is because from triangle inequality in $H$ we have (please note that there is no longer an equality in the left part because $w$ is not on $P_H$ and we need paths for $d_H(u, w), d_H(w, v)$ that contain some edges from outside of $P_H$, but the equality in the right part which we need is still true as $w \in P_G$):

$$d_H(u, w) + d_H(w, v) \geq d_H(u, v) > td_G(u, v) = td_G(u, w) + td_G(w, v)$$

3

then by same argument as previously, either $d_H(u, w) > td_G(u, w)$ or $d_H(w, v) > td_G(w, v)$. By substituting $u \leftarrow w$ as many times as needed, we finally have that $u, v$ are neighbours in $G$ (please note that by such substitution, we can again have intersections between $P_G, P_H$ and we may need to move to the previous step, but $|P_G|$ constantly decreases).

- As we now have a counterexample where $|P_G| = 1$ and $|P_H| > t|P_G| = t$, we arrive at a contradiction: when we were processing the only edge on $P_G$, we should have added it to $H$ as it didn't cause a cycle of length $\leq t + 1$ - it doesn't cause that now as $d_H(u, v) = |P_H| \geq t + 1$ and in the time of processing of that edge it also didn't as $H$ at that time was a subgraph of the final $H$.

$\square$

2. The graph is sparse enough:

- Let $d = 2m/n$ be average degree.
- There is subgraph $H$ of $G$ with minimum degree $d' = d/2$: keep removing vertices with degree smaller than $d'$. We cannot end with no vertices, since then we removed less than $m$ edges.
- $H$ has every vertex of degree at least $m/n$, and no cycle of length $t + 1$ or less.
- BFS tree of depth $t/2$ has no cycles in $H$
- $(m/n - 1)^{t/2} \leq |H| \leq n$, which implies bound on $m$

$\square$

# References

[Che09]  Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009.

[FWY19]  Manuel Fernandez, David P. Woodruff, and Taisuke Yasuda. Graph spanners in the message-passing model. *CoRR*, abs/1911.05991, 2019.