Smaller Coresets For k-Median and k-Means Clustering

Har-Peled and Kushal
SoCG 2005

Sergei Vassilvitskii
March 15, 2006

Problem Recap

Given: point set P of n points in $\mathcal{R}^d$.

Find: A weighted set C of k points, such that:

`k-median`: $\nu_C = \min \sum_{p \in P} w(p) d(p, C)$

`k-means`: $\mu_c = \min \sum_{p \in P} w(p) d(p, C)^2$

A weighted set $\mathcal{S}$ is a (k, ϵ)- coreset, if for any set C on k points:

$$(1 - \epsilon)\nu_C(P) \leq \nu_C(\mathcal{S}) \leq (1 + \epsilon)\nu_C(P)$$

Results

Last time showed how to find (k, ϵ)-coresets of size $O(k\epsilon^{-d} \log n)$ for both k-means and k-median clustering.

This yielded a k-median algorithm with running time:

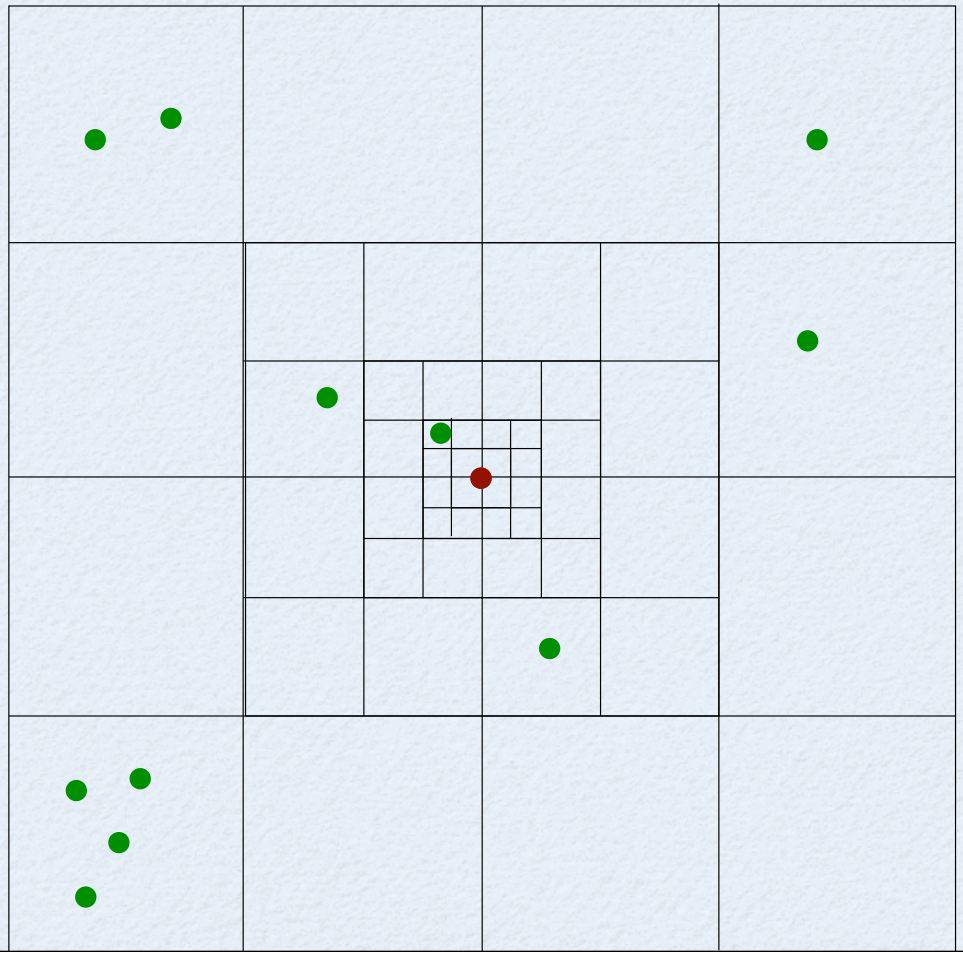
$$O\left(n + e^{O\left(\frac{1+\log(1/\epsilon)}{\epsilon}\right)^d} k^{O(1)} \log^{O(1)} n\right)$$

Today: Improve the coreset size to $O(k^2/\epsilon^d)$ for k-median, $O(k^3/\epsilon^{d+1})$ for k-means.

Unfortunately, no improvement in overall algorithm run time.

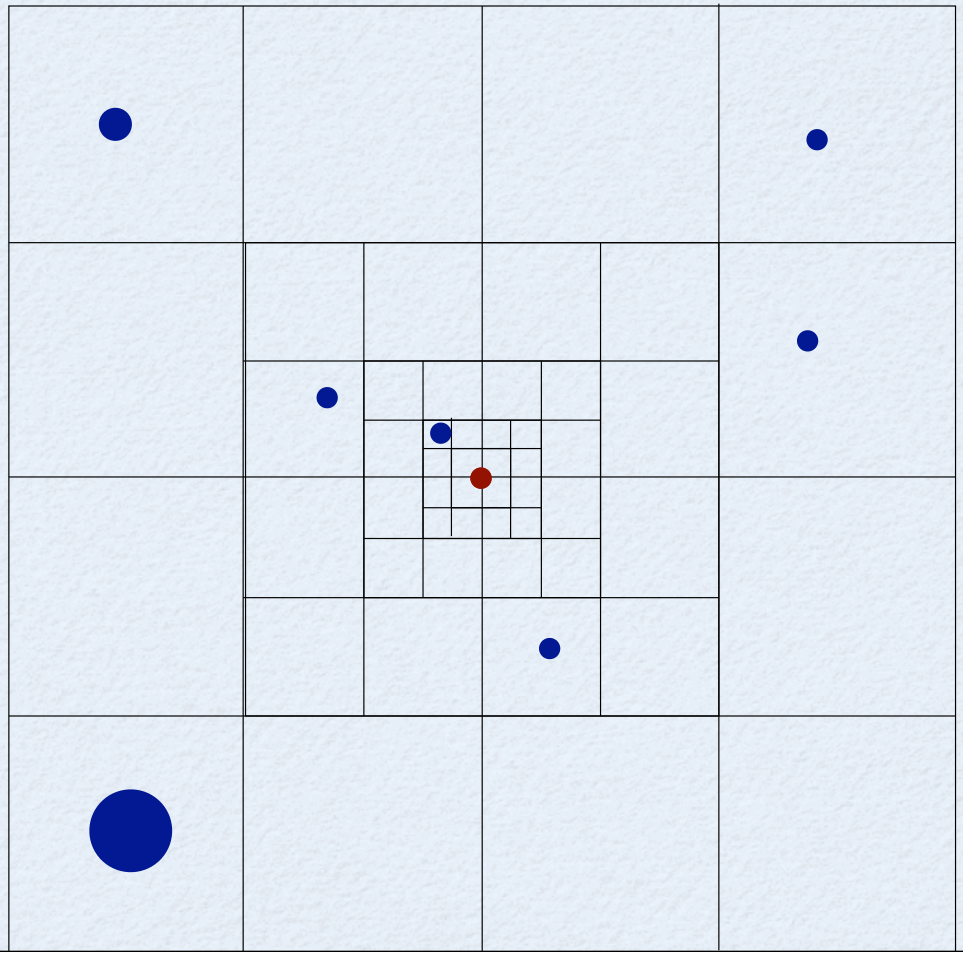
Main Idea

Last time: Found a constant factor approximation, and snapped points to an exponential grid around those points.



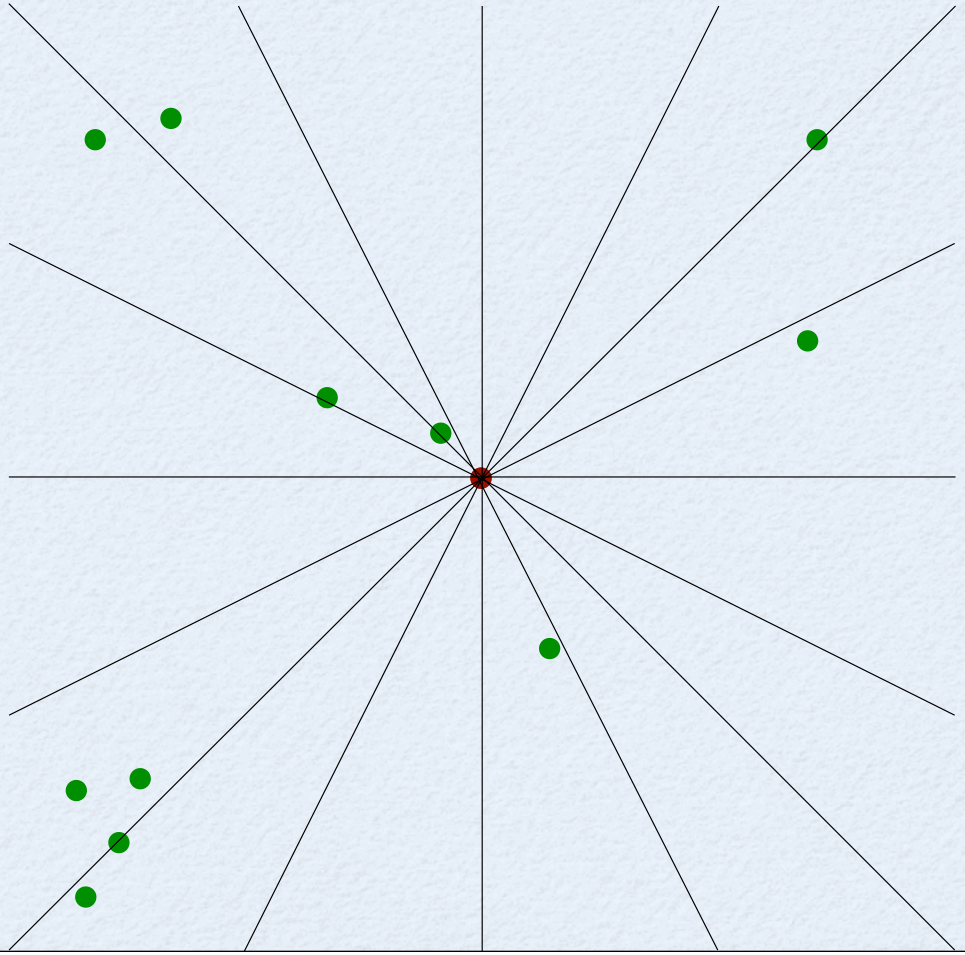
Main Idea

Last time: Found a constant factor approximation, and snapped points to an exponential grid around those points.



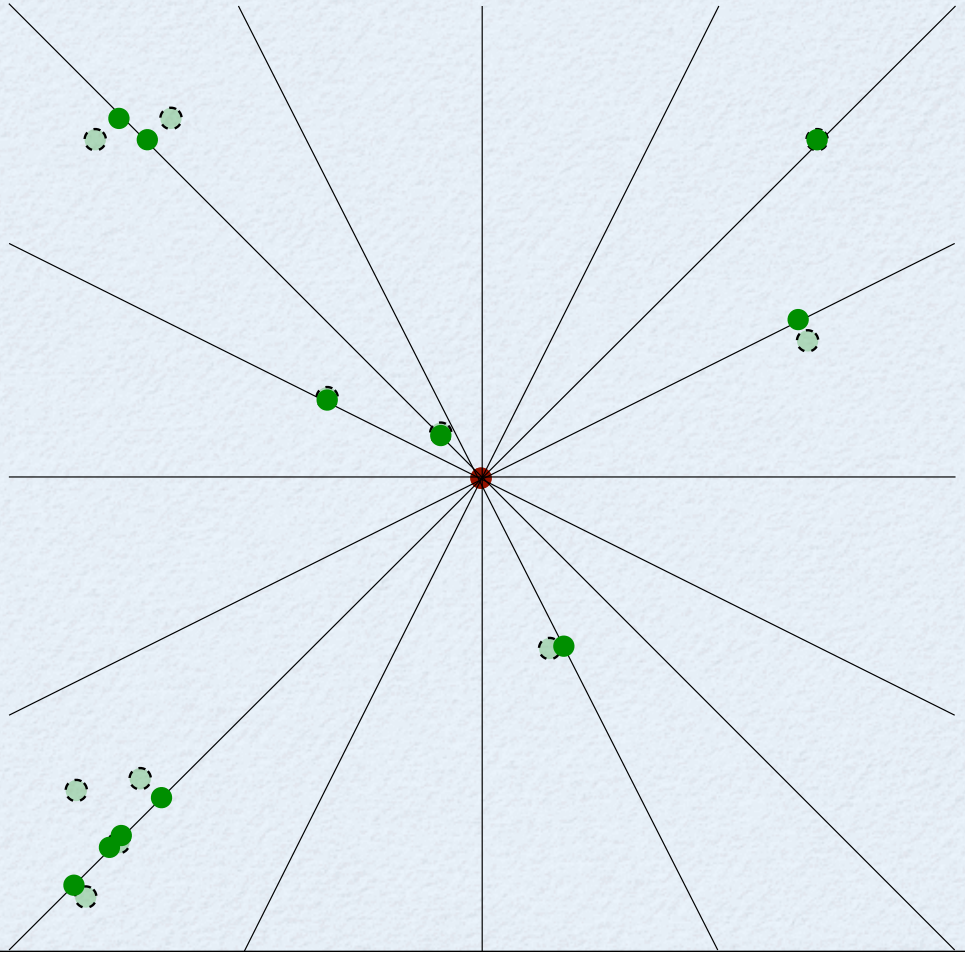
New Idea

Instead of grid, snap to lines:



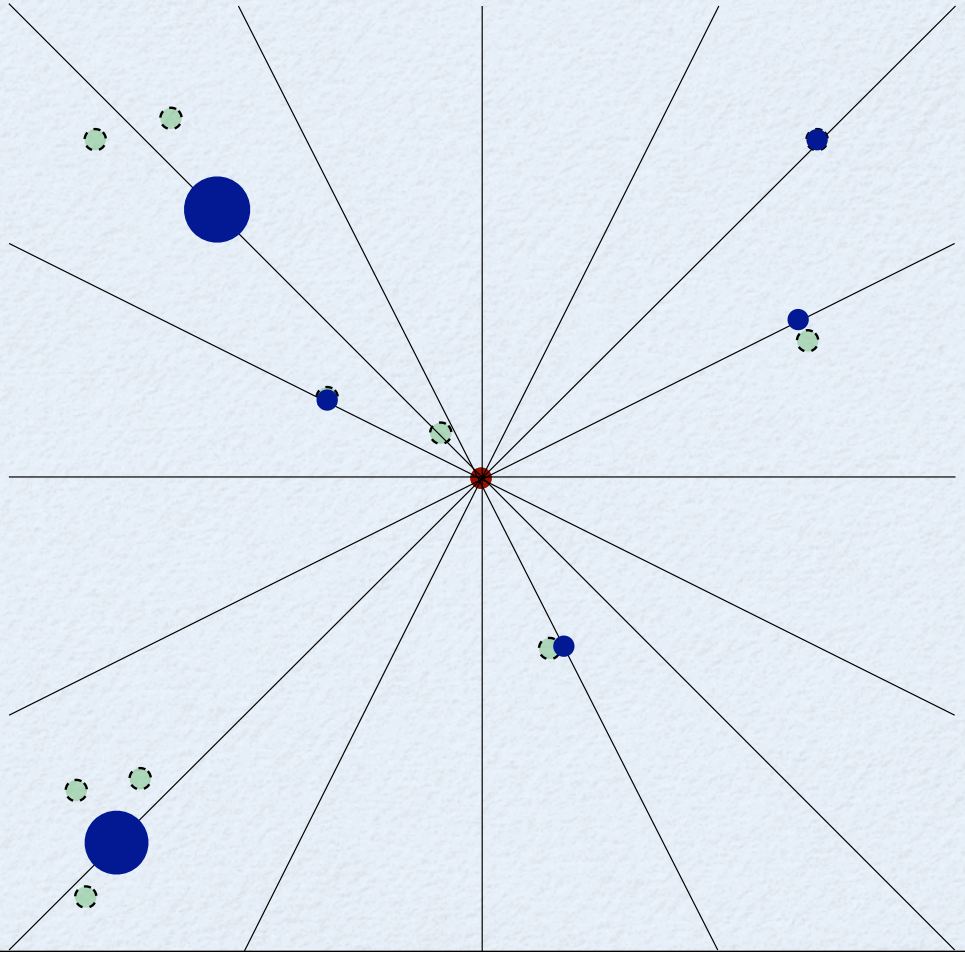
New Idea

Instead of grid, snap to lines:



New Idea

Compute a coreset for each line.



Proving the Results

Proof Outline:

1. Prove some basic facts

2. Find a coreset on a line

3. Intuition

4. Technical Lemma (Proof for higher d)

5. Tying up loose ends

Notation

Notation:

For a pointset P , let $\bar{m}(P) = \frac{1}{w(P)} \sum_{p \in P} w(p)p$ be the mean. (Often will just say $\bar{m}$)

Define the cumulative error:

$$\mathcal{E}_\nu(P) = \nu_{\bar{m}}(P) = \sum_{p \in P} w_p \|p\bar{m}\|$$

For a set of points B in $\mathcal{R}$, let $\mathcal{I}(B)$ be the smallest interval containing B .

Basic Facts

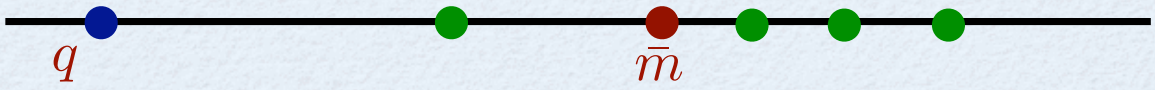
Let P be a set of weighted points on a line ℓ .

If L and R are the points of P left and right of $\bar{m}$, then:

$$\sum_L w_p \|\bar{m}p\| = \sum_R w_p \|\bar{m}p\|$$

For $q \in \ell, q \notin \mathcal{I}(P)$ $\quad \nu_q(P) = w(P)\|q\bar{m}\|$

Proof:



Basic Facts (2)

For any set of points $C \subseteq \mathcal{R}^d$, if $\mathcal{S}$ is a coreset made of a single point $\bar{m}$ with weight $w(P)$, then $|\nu_C(P) - \nu_C(\mathcal{S})| \leq \mathcal{E}_\nu(P)$.

Proof: $|\nu_C(P) - \nu_C(\mathcal{S})| = \left| \sum_{p \in P} w_p (d(p, C) - d(\bar{m}, C)) \right|$

$$\leq \sum_{p \in P} w_p \|p\bar{m}\| = \mathcal{E}_\nu(P)$$

For any $P \subseteq \mathcal{R}$, we have $\mathcal{E}_\nu(P) \leq 2\nu_{opt}(P)$ where $\nu_{opt}(P)$ is the optimal one clustering.

Coreset on a Line

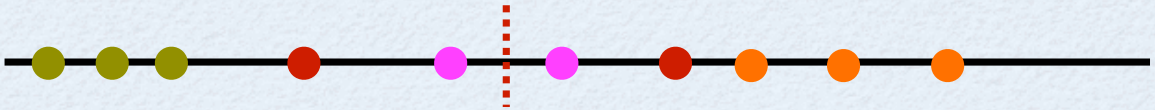
Let V^* be a c approximation to value of optimal solution. Scan line from left to right, group points into batches with error less than: $\phi = \frac{\epsilon}{10ck} V^*$

Take the mean of each batch to create a coreset.

Coreset Correctness

What is the number of batches? Each has error $\phi = \frac{\epsilon}{10ck} V^*$. Easy to verify, no more than $O(k/\epsilon)$ batches.

Consider any set of clusters C .



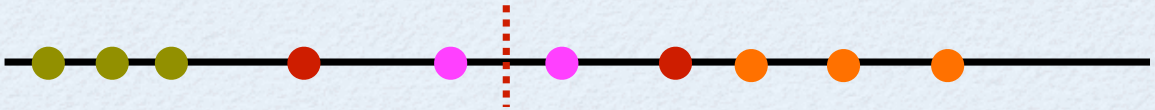
If a batch B is fully covered by one cluster, and the cluster c_i is outside the batch, then: $\nu_{c_i}(B) = w(B) \cdot \|\bar{m}(B)x_i\|$.

The contribution is identical for the coreset.

Coreset Correctness

What is the number of batches? Each has error $\phi = \frac{\epsilon}{10ck} V^*$. Easy to verify, no more than $O(k/\epsilon)$ batches.

Consider any set of clusters C .



The batches contributing to the error have a Voronoi boundary, or contain a cluster center on the interior.

These contribute at most $\mathcal{E}_\nu(B)$ to the error.

Total error: $\sum_{i=1}^{2k-1} \mathcal{E}_\nu(B) \leq (2k-1)\phi \leq \epsilon \nu_{opt}$

Proving the Results

Proof Outline:

~~1. Prove some basic facts~~

~~2. Find a coreset on a line~~

3. Intuition

4. Technical Lemma (Proof for higher d)

5. Tying up loose ends

Coreset Construction

1. Find C, a c approximation to the optimal solution.

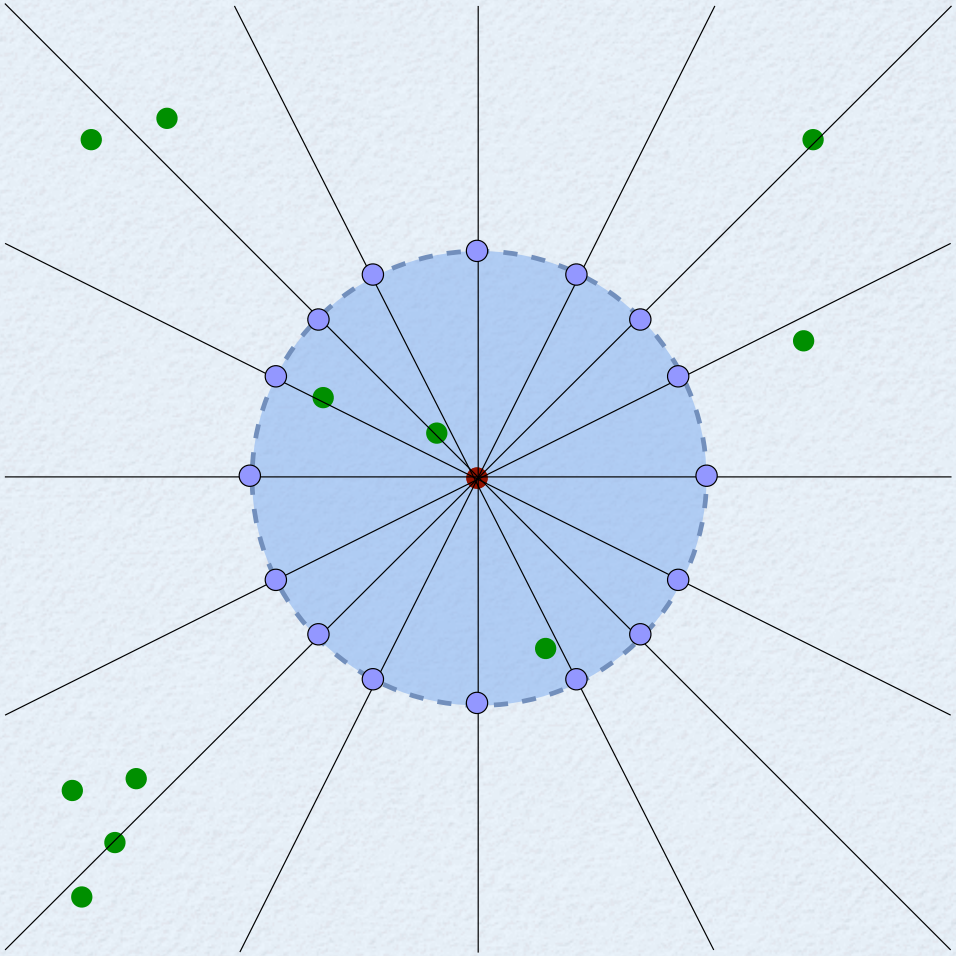
2. Find Q, an $\epsilon/(3c)$ coreset for a unit sphere centered at each center in C.

3. Generate a fan of lines around each center in C through every point in Q.

4. Snap points to closest line.

5. Compute a coreset for each line by grouping points in batches as before.

Coreset Construction



Coreset Construction

1. Find C, a c approximation to the optimal solution.

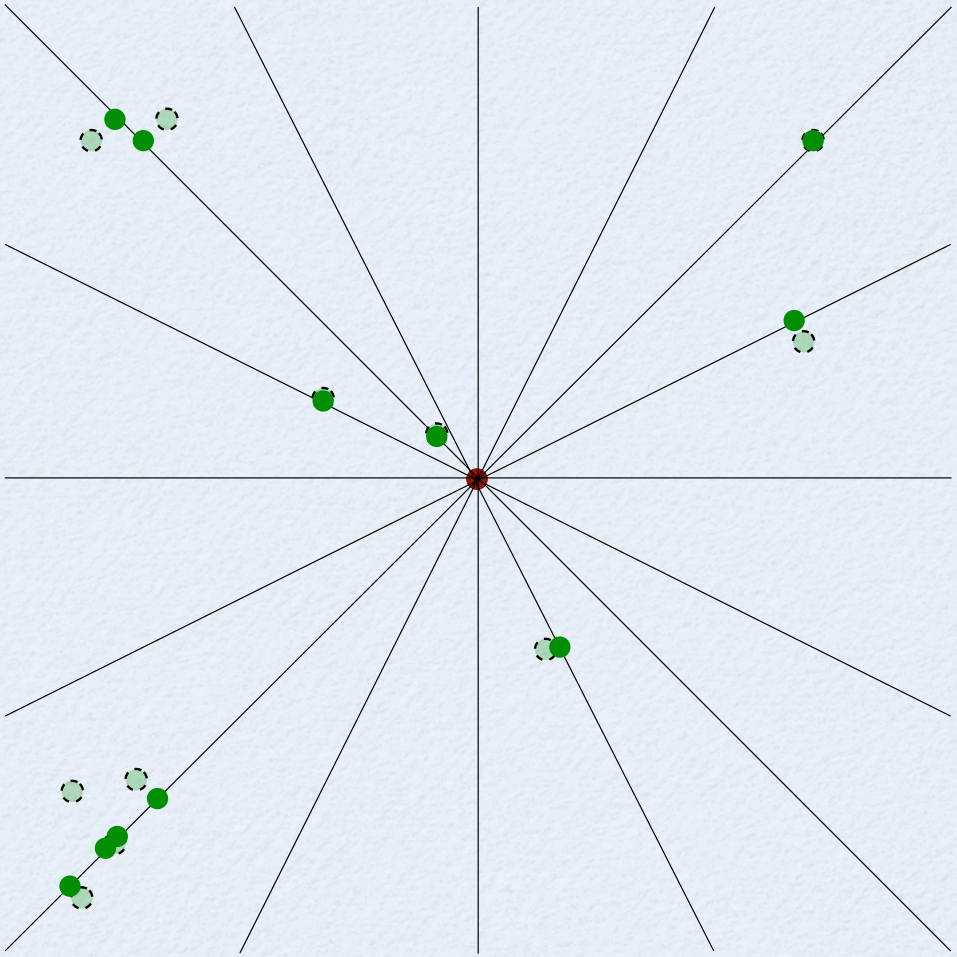
2. Find Q, an $\epsilon/(3c)$ coreset for a unit sphere centered at each center in C.

3. Generate a fan of lines around each center in C through every point in Q.

4. Snap points to closest line.

5. Compute a coreset for each line by grouping points in batches as before.

Coreset Construction



Coreset Construction

1. Find C, a c approximation to the optimal solution.

2. Find Q, an $\epsilon/(3c)$ coreset for a unit sphere centered at each center in C.

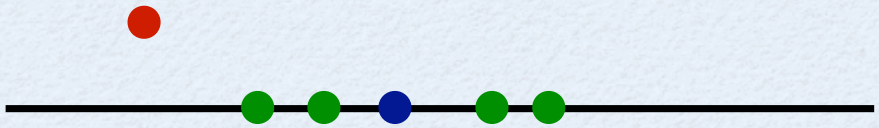
3. Generate a fan of lines around each center in C through every point in Q.

4. Snap points to closest line.

5. Compute a coreset for each line by grouping points in batches as before.

Intuition

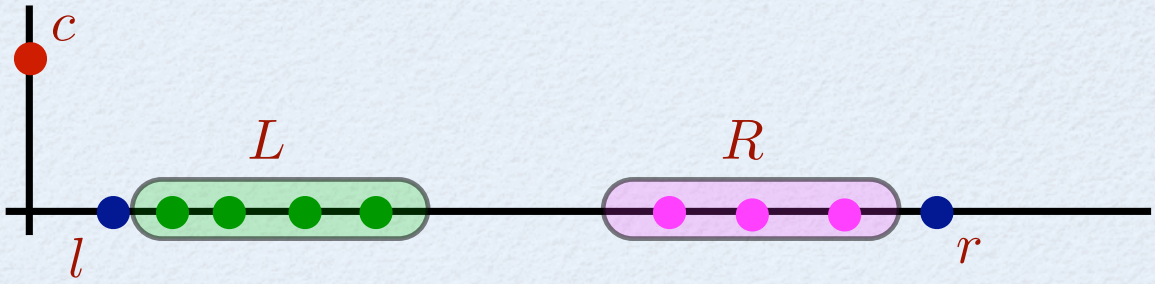
We need cancellation to occur:



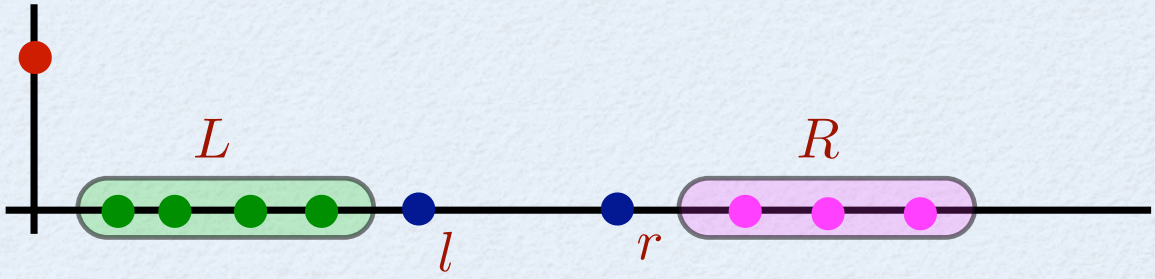
Replacing ● ● ● ● by ● increases some distance to the optimal center, but decreases others. Need to carefully count this cancellation.

Technical Lemma

$$\mathcal{S}_L = (l, w(L)) \quad \mathcal{S}_R = (r, w(R))$$
$$\nu_l(L) = \nu_r(R)$$

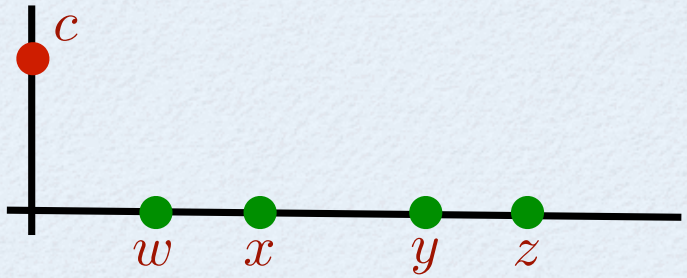


Then: $E = \nu_c(L) + \nu_c(R) - \nu_C(\mathcal{S}_L) - \nu_C(\mathcal{S}_R) \leq 0$



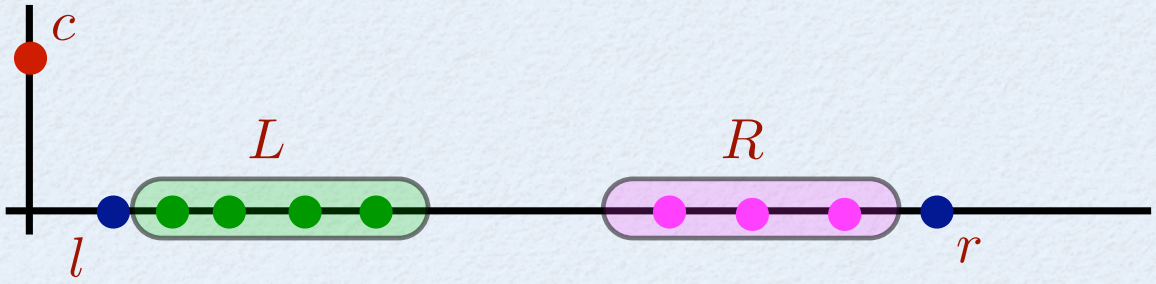
$$E = \nu_c(L) + \nu_c(R) - \nu_C(\mathcal{S}_L) - \nu_C(\mathcal{S}_R) \geq 0$$

Proof



$$e(x,y) = \|yc\| - \|xc\|$$

$$\frac{e(w,x)}{\|wx\|} \leq \frac{e(y,z)}{\|yz\|}$$

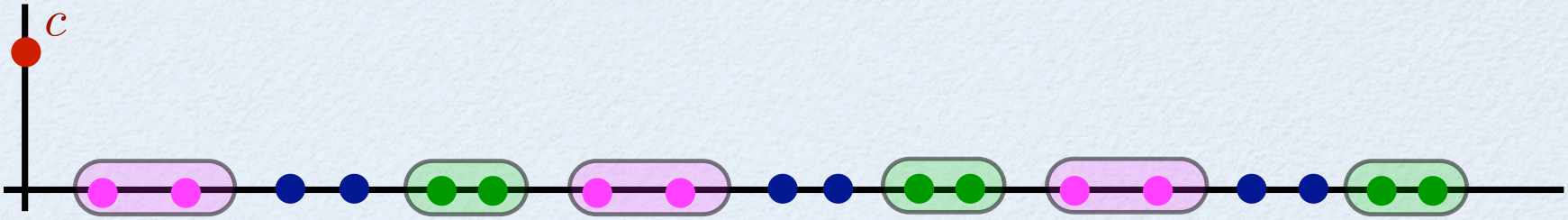


$$E = \nu_c(L)+\nu_c(R)-\nu_C(\mathcal{S}_L)-\nu_C(\mathcal{S}_R) = \sum_{p\in L} w_p(\|cp\|-\|cl\|)+\sum_{p\in R} w_p(\|cp\|-\|cr\|)$$

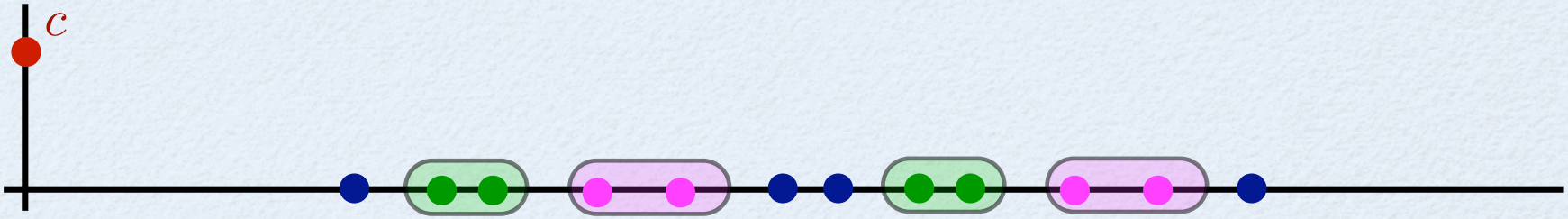
$$= \sum_{p\in L} w_p e(l,p) - \sum_{p\in R} w_p e(p,r) \quad = \sum_{p\in L} w_p \|pl\| \frac{e(l,p)}{\|pl\|} - \sum_{p\in R} w_p \|pr\| \frac{e(p,r)}{\|pr\|}$$

$$\leq \sum_{p\in L} w_p \|pl\| z - \sum_{p\in R} w_p \|rp\| z \quad = z(\nu_l(L) - \nu_r(R)) = 0$$

Cancellation



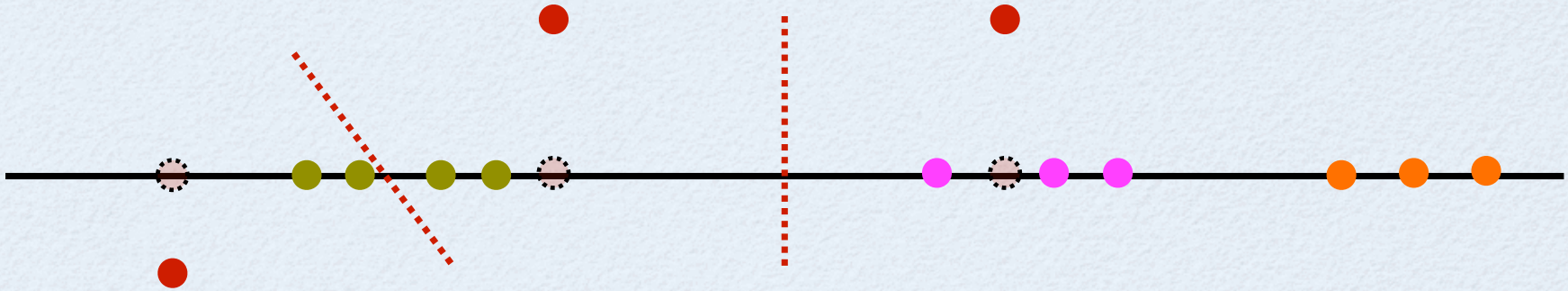
Assignment 1: Total Error positive.



But, this subset has negative total error.

Therefore, total error bounded by error of two extremal sets

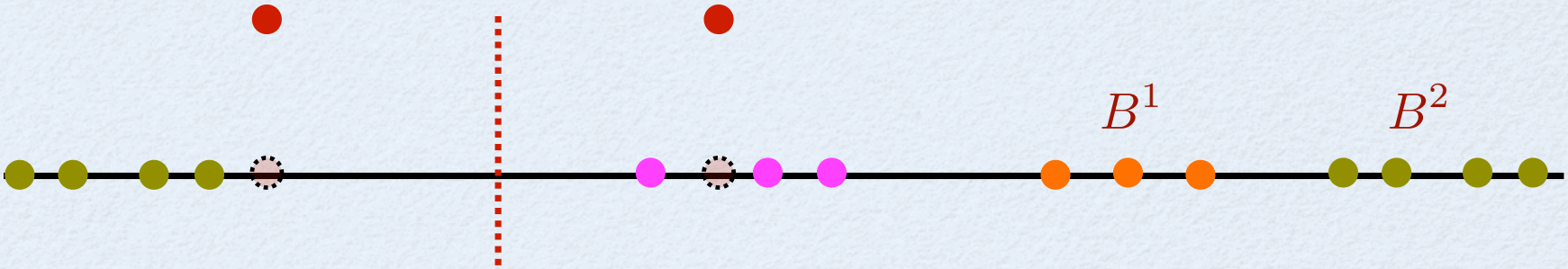
Coreset Proof



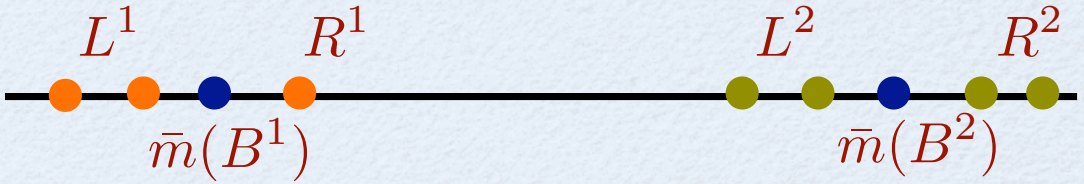
Consider batches served by two or more centers, and those that contain the projection of the center. These contribute at most:

$$\sum_B \mathcal{E}_\nu(B) \leq 2k \cdot O(\frac{\epsilon}{k})\nu_{opt}(P) = O(\epsilon)\nu_{opt}(P)$$

Proof (2)

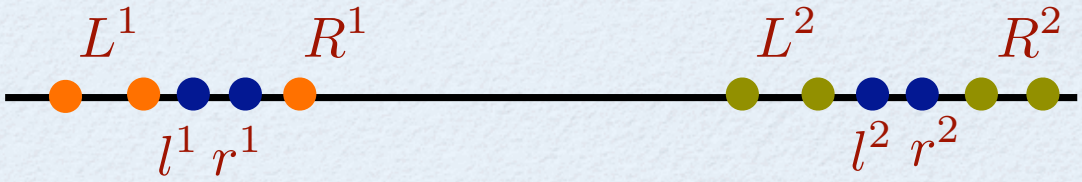


Coreset for each batch is the mean point.

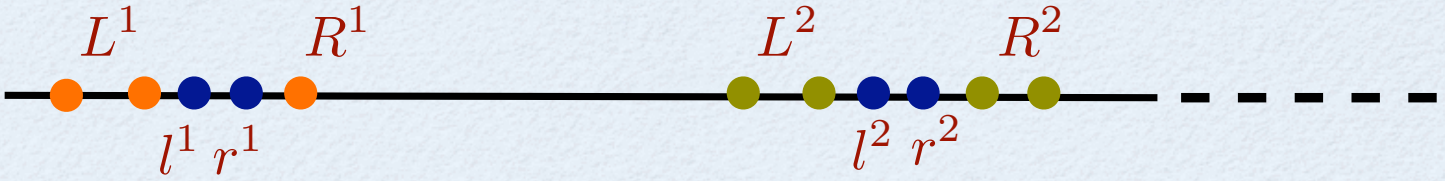


Each batch has same error! (by construction)

Split each mean point into two.



Proof (3)



$$E = \sum_{i=1}^{t} \left(\nu_c(L^i) - \nu_c(\mathcal{S}_l^i) + \nu_c(R^i) - \nu_c(\mathcal{S}_r^i) \right)$$

$$E' = \sum_{i=1}^{t-1} \left(\nu_c(R^i) - \nu_c(\mathcal{S}_r^i) + \nu_c(L^i) - \nu_c(\mathcal{S}_l^i) \right) \leq 0$$

$$E = E' + \nu_c(L^1) - \nu_c(\mathcal{S}_l^1) + \nu_c(R^t) - \nu_c(\mathcal{S}_r^t)$$

$$\leq \mathcal{E}_\nu(B^1) + \mathcal{E}_\nu(B^t) = O(\epsilon/k)\nu'_{opt}$$

Therefore each line contributes a total of $O(\epsilon)\nu_{opt}(\ell)$

Proving the Results

Proof Outline:

~~1. Prove some basic facts~~

~~2. Find a coreset on a line~~

~~3. Intuition~~

~~4. Technical Lemma (Proof for higher d)~~

5. Tying up loose ends

Finishing up

For a point p, let p' be the snapped point. Let c_i be the corresponding c approximate center. Then:

$$\|pp'\| \leq \|pc_i\|\epsilon/(3c)$$

The total snapping error is:

$$\sum_{p\in P} \|pp'\| \leq \sum_{i=1}^{k} \sum_{p\in P_i} (\epsilon/3c)\|pc_i\| \leq \epsilon/3\nu_{opt}$$

Total coreset error:

$$|\nu_c P - \nu_C(\mathcal{S})| \leq |\nu_C(P) - \nu_C(P')| + |\nu_C(P') - \nu_C(\mathcal{S})|$$

$$\leq (\epsilon/3)\nu_{opt} + O(\epsilon)\nu_{opt} = O(\epsilon)\nu_{opt}$$

Conclusions

While we found a smaller coreset, the running time is virtually unchanged. The dominating factor in the running time is solving the small instance.

1. Can we improve the exact algorithm on the coreset?

2. Can we find coresets with no dependency on n and only a polynomial dependency on the dimension?

Thank You