Next part of the course: Streaming algorithms for metric problems & graph problems.

Today: Clustering

We'll see: 8-approx alg for k-center

$2+O(\varepsilon)$-approx for k-center

Data stream: sequence of tokens $\sigma_1 \sigma_2 \dots \sigma_m$ from $[n]$ or $\mathbb{R}^n$ or edges of G, etc.

Each elt is inserted in the stream sequentially

Alg can store some info about stream seen so far.

Alg cannot access elts not saved that appeared earlier (unless it makes more passes over the data)

Goal: • Approximately solve an optimization problem

• Use sublinear space in m, n

Ideally : $O(\log m, \log n)$

↑ space to output a constant number of elts of $[n]$.

Sometimes cannot do better than $\Omega(m)$

(we'll see lower bounds later in the course)

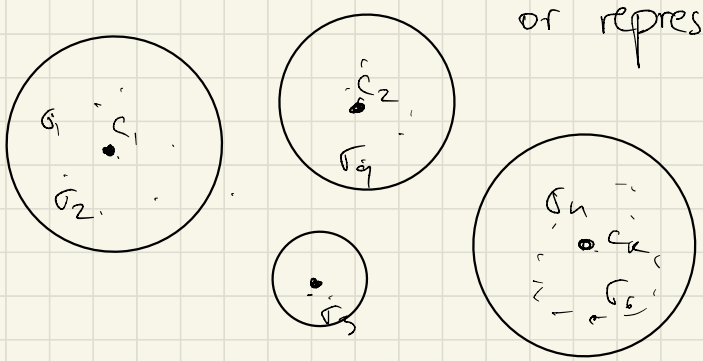
Today: clustering in metric spaces.

Given a data stream of pts in $\mathbb{R}^n$

$\sigma = \sigma_1 \sigma_2 \dots \sigma_m$ $\qquad \sigma_i \in \mathbb{R}^n$

output a good "summary" of σ

Reasonable summary : centers of k clusters or representatives of each cluster.



$c_1, c_2, \dots c_k$.

Algs in this lecture work for more general metrics.

Metric space: (M, d) $\quad d : M \times M \to \mathbb{R}^+$

- $d(x,y) = 0$ iff $x = y$
- $d(x,y) = d(y,x)$
- $d(x,y) \leq d(x,z) + d(z,y)$

Eg: $\mathbb{R}^n$ &

$d(x,y) = \|x-y\|_p \qquad p > 0.$

$p = 2$ Euclidean distance.

$p = \infty \qquad \max |x_i - y_i|$

For graphs $M = V$ (vtx set)

$d(x,y) =$ length of shortest path

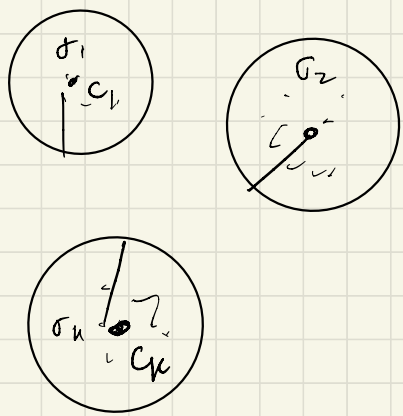
Clustering

Extend $d(x,y)$ to $d(x,S) = \min_{y \in S} d(x,y)$

Standard objectives : minimize

- $\Delta_\infty(\sigma, R) = \max_{x \in \sigma} d(x, R)$

(k-center)



We'll focus on k-center in this lecture.

Other cost measures:

$$\Delta_1(\sigma, R) = \sum_{x \in \sigma} d(x, R)$$

(k-median)

$$\Delta_2(\sigma, R) = \sum_{x \in \sigma} d(x, R)^2$$

(k-means)

Goal: Choose $|R| \leq k$ centers to min cost $\Delta(\sigma, R)$

Hochbaum & Shmoys (Doubling Alg)

Init: $R \leftarrow$ set of first $k+1$ tokens
(centers) $(\sigma_1, \sigma_2 \ldots \sigma_{k+1})$

$\tau =$ dist bet closest pair (z, y)

$R \leftarrow R \setminus \{z\}$

We'll forget abt space needed to repres. a token

on new token x

- if $\min_{r \in R} d(x, r) \geq 2\tau$

if cost is high double the threshold.

$R \leftarrow R \cup \{x\}$

while $|R| > k$

$R \leftarrow$ maximal set $R' \subset R$

st $\forall r \neq s \in R'$

$d(r, s) \geq \tau$.

$\tau \leftarrow 2\tau$.

At end of stream output R.

Analysis :

Invariants :

- separation : $\forall r \neq s \in R \quad d(r,s) \geq \tau$
- cost : $\Delta_\infty(\sigma, R) \leq 2\tau$

Verifying the inv.

inv 1. holds after init phase ✓

assume it holds before processing x

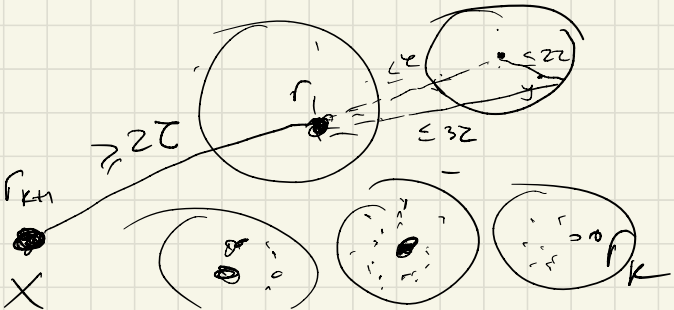
we'll show it hold after processing x

By design ✓

inv 2: in initializ phase $\Delta_\infty(\sigma, R) = \tau$

$< 2\tau$ ✓

Assume inv 2 holds before processing x

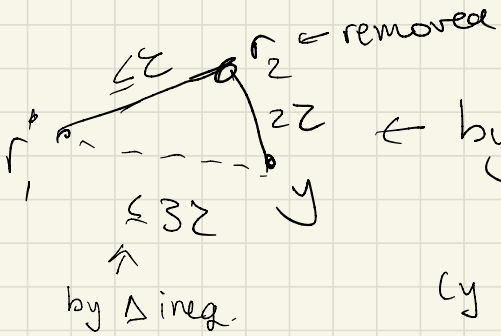


$\tau' = 2\tau.$

$\Delta(\sigma, R') \text{ vs } 4\tau.$

- what dist. an r that is kicked out of R is from R' : τ.

- what dist an elt y in the cluster of r (that was removed) is from R'



← by assumpt before processing x (y was represented by r_2)

$$3\tau < 4\tau = 2\tau' \quad ✓$$

Thm: Alg above use $O(k)$ space & produces an 8-approximation to opt clustering.

Pf: Let R^* is an OPT clustering producing opt cost $\Delta_\infty(\sigma, R^*)$.

Let $\hat{R}$, $\hat{\tau}$ the final output

$d(x, y) \geq \hat{\tau}/2$ before final update to $\hat{\tau}$

$\Delta_\infty(\sigma, \hat{R}) \leq 2\hat{\tau}$ (by inv. 2).

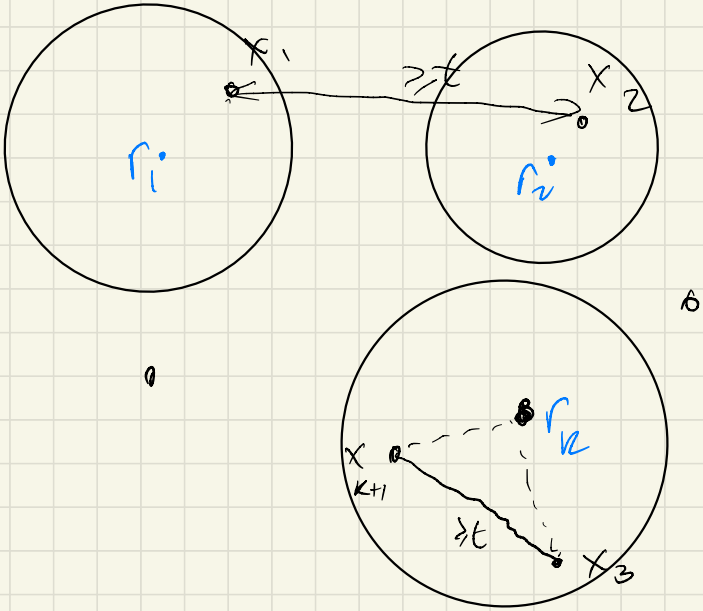
How to relate $\Delta_\infty(\sigma, R^*)$ to $\hat{\tau}$?

<u>Lemma</u>: Given $x_1, x_2 \ldots x_{k+1} \in \sigma$

st $d(x_i, x_j) \geq t$ $\forall i, j \in [k+1]$.

Then $\forall R \subseteq M$ of $|R| \leq k$

$$\Delta_\infty(\sigma, R) \geq t/2$$



$\exists$ 2 x_i's represented by same r_i

$$2 \max_k d(r_k, x_j) \geq d(x_{k+1}, r_k) + d(r_k, x_3) \geq t$$

$$\leq 2 \cdot \Delta_\infty(\sigma, R)$$

$$\Rightarrow \Delta_\infty(\sigma, R) \geq t/2$$

any R, so OPT too

Applying Lemma to alg analysis:

Lem is true for opt R^*, so

$$\Delta_\infty(\sigma, R^*) \geq t/2$$

Recall every $r_i, r_j \in R'$ are st

$$\text{dist}(r_i, r_j) \geq \hat{\varepsilon}/2.$$

$$\Rightarrow \Delta_\infty(\sigma, R^*) \geq \hat{\varepsilon}/4$$

$\uparrow$ OPT

Recall output $\hat{R}$ st. $\Delta_\infty(\sigma, \hat{R}) \leq 2\hat{\varepsilon}$

$$\Rightarrow \Delta_\infty(\sigma, \hat{R}) \leq 8\, \Delta_\infty(\sigma, R^*)$$

$\uparrow$ OPT.

Metric property of cost fnct:

$\Delta(\sigma, R)$ satisf. metric prop if $\forall\, \underbrace{\sigma, \bar{\pi}}_{\text{streams}}$

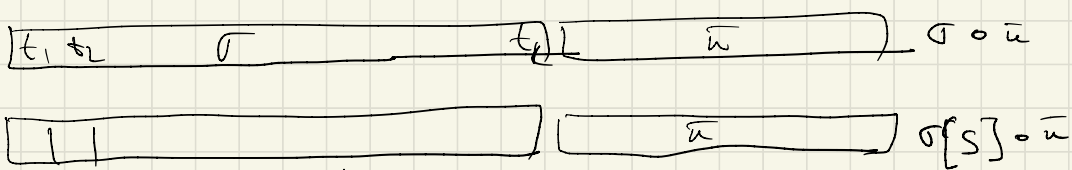
& $\forall$ summary S for σ, $\forall$ summary T for $\sigma \circ \bar{\pi}$

$$\Delta(\sigma[S] \circ \bar{\pi}, T) - \Delta(\sigma, S) \leq$$
$$\leq \Delta(\sigma \circ \bar{\pi}, T)$$
$$\leq \Delta(\sigma[S] \circ \bar{\pi}, T) + \Delta(\sigma, S)$$

Above

S is a sumary for σ

T is a summary for $\sigma \circ \bar{\pi}$ or for $\sigma[S] \circ \bar{\pi}$



$rep(t_1)$ $\quad$ $rep(t_k)$ $\quad$ replace each t_i with its representative in S.

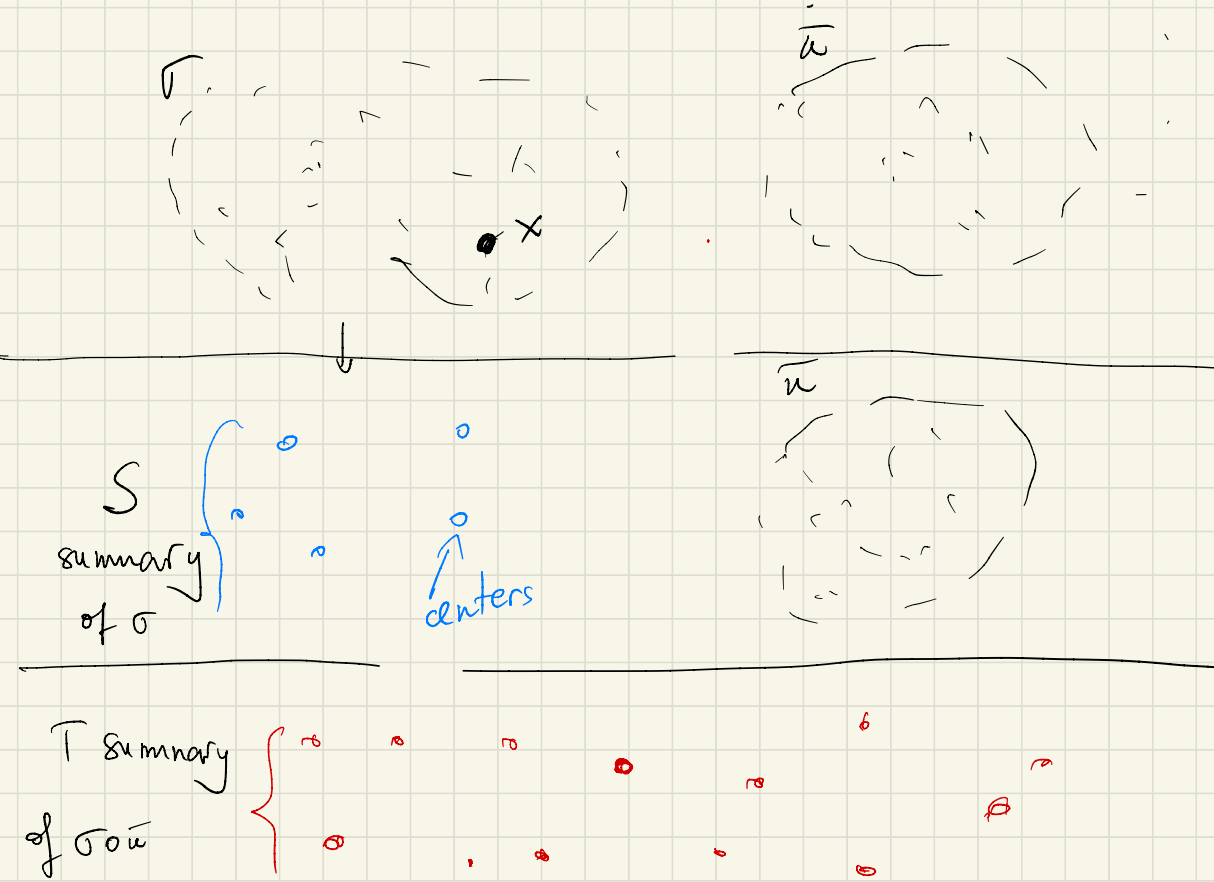
closest centers

& continue reading $\bar{\pi}$

Ineq above says that if S is a good summary for σ & T is a good summary for $\sigma \circ \bar{\pi}$ then T is still a good summary for $S \circ \bar{\pi}$; so we won't make much error on forgetting σ.

Claim: $\Delta_\infty(\sigma, R)$ is a metric cost measure.

Pf idea: Prove defining ineq holds for every element of $\sigma \circ \bar{\pi}$.



How does x, $rep_S(x)$, $rep_T(x)$ compare?

$d(x, S)$ vs $d(x, T)$ vs $d(S, T)$

(homework)

Def: Let Δ be an error cost fnc

& $\alpha \geq 1 \in \mathbb{R}$.

An α-threshold alg for Δ:

on input threshold t and stream σ

Either ①: Alg produces a summary S

with $\Delta(\sigma, S) \leq \alpha \cdot t$

OR

②. Alg fails

if no soln T exists of

error cost $\Delta(\sigma, T) < t$

(i.e. so for OPT T^* we

must have $\Delta(\sigma, T^*) > t$)

Claim:

Doubling alg above is a 2-threshold alg.

Pf: Given thresh τ

either · outputs sol of cost $\leq 2\tau$

or · fails (no sol of cost $< \tau$ exists)

Thm: If Δ is a metric cost

& has an α-approxim. thresh alg

Then $\exists$ an $\alpha + O(\varepsilon)$ - approx alg.

Corollary: For Δ_∞ we have a $2 + O(\varepsilon)$ - approx for k-center.

in space is indep. of $|\sigma|$, val OPT

Proof of Theorem (*)

Guha's cascading alg:

- Read 1st B items of input while sum. cost $= 0$ $(S_\infty = 0)$ till instance of cost > 0. Let c be this cost.

- Run an α-threshold alg in parallel for thresholds

$$\left.\begin{array}{l} c \\ c(1+\varepsilon) \\ \vdots \\ c(1+\varepsilon)^J \end{array}\right\} \frac{1}{\varepsilon}\log\frac{\alpha}{\varepsilon} \text{ many times}$$

each machine operates on small space

s.t. $(1+\varepsilon)^J \geq \frac{\alpha}{\varepsilon}$

- If some thread fails (no sol of cost $\Delta_\infty < t = c(1+\varepsilon)^i$ exists) ← by def of α-thresh alg.

Then Update all thresholds $< t$ to

$$(1+\varepsilon)^J \cdot t = (1+\varepsilon)^J c \cdot (1+\varepsilon)$$

$$\approx \alpha/\varepsilon$$

- Start processing remaining stream from current summarization. (maintain same # of copies)
- Repeat until end of input.

Output sumary of an instance that is still running, with smallest threshold.

Lecture 16

Thm: Cascading alg is an $\alpha + O(\varepsilon)$-approx if run on an α-thresholding alg, for metric cost Δ, in space $O\left(\frac{1}{\varepsilon} \log \alpha/\varepsilon \cdot (\text{space of } \alpha\text{-thresh alg})\right)$

Corollary: For Δ_∞ & k-center gives $2+O(\varepsilon)$-approx in space $O\left(\frac{k}{\varepsilon} \log \frac{1}{\varepsilon}\right)$

Proof of Thm:

Let t be the threshold used by the instance with final output. Instance has lowest current threshold.

How many times this threshold changed?

Say j times. The thresholds on that thread are:

$$\underset{\substack{\uparrow \\ \text{init} \\ = C\cdot(1+\varepsilon)^i \\ \text{for some } i}}{t_0} \qquad \underbrace{t_0(1+\varepsilon)^J}_{t_1} \quad \cdots \quad \underbrace{t_0(1+\varepsilon)^{J\cdot j}}_{t_j = T}$$

So $t_i = \dfrac{t}{(1+\varepsilon)^{(j-i)J}}$

Consider stream σ_i to be the stream occuring bet the $i-1$ & ith change of thresh.

$$\sigma = \sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_{j+1}$$

Let S_1 S_2 S_{j+1} be the summaries before each jump.

Note: S_i is a summary of $S_{i-1} \circ \sigma_i$ (not of σ_i)

$S_0 = \phi$.

Claim: $\forall i$

$$\Delta(\sigma_1 \circ \sigma_2 \circ \cdots \sigma_i, S_i) \leq \Delta(S_{i-1} \circ \sigma_i, S_i) + \Delta(\sigma_1 \circ \cdots \sigma_{i-1}, S_{i-1})$$

Proof: $i=1$ $\quad \Delta(\sigma_1, S_1) \le \Delta(\sigma_1, S_1) + \Delta(\sigma_0, S_0)$

$$\cancel{\Delta(\sigma_1 \circ \sigma_2, S_2)} \le \Delta(S_1 \circ \sigma_2, S_2) + \Delta(\sigma_1, S_1)$$

cancel out

(by metric property)

$$\Delta(\sigma_1 \circ \sigma_2 \circ \sigma_3, S_3) \le \cancel{\Delta(\sigma_1 \circ \sigma_2, S_2)} + \Delta(S_2 \circ \sigma_3, S_3)$$

$\vdots$

$$\Delta(\sigma_1 \circ \sigma_2 \circ \dots \circ \sigma_{j+1}, S_{j+1}) \le \cancel{\Delta(\sigma_1 \circ \sigma_2 \dots \circ \sigma_j, S_j)} + \Delta(S_j \circ \sigma_{j+1}, S_{j+1})$$

$$\Delta(\sigma_1 \circ \sigma_2 \circ \dots \circ \sigma_{j+1}, S_{j+1}) \le \sum_{i=0}^{j+1} \Delta(S_{i-1} \circ \sigma_i, S_i)$$

$\star$

we'll bound each term.

Recall that since summary S_i is output by an α-thresholding alg for thresh. $t_{i-1} = \frac{T}{(1+\varepsilon)^{J\cdot(j-i+1)}}$

we must have output a sol. of cost

$$\Delta(S_{i-1} \circ \sigma_i, S_i) \leq \alpha \cdot t_{i-1}$$

$$= \alpha \cdot \frac{t}{(1+\varepsilon)^{J(j-i+1)}}$$

So $\circledast \leq \sum_{i=1}^{j+1} \alpha \cdot \frac{t}{(1+\varepsilon)^{J(j-i+1)}} =$

$$= \alpha \cdot t \sum_i \left(\frac{1}{(1+\varepsilon)^J}\right)^{j-i+1}$$

$$\leq (\alpha + O(\varepsilon))\, t \qquad (**)$$

How does t compare to OPT ?

Recall t is last threshold that didn't cause any thread to fail.

So some thread must have failed on

$$\frac{t}{(1+\varepsilon)}$$

by prop. of a failing α-thresh. alg.

So $$\Delta(\sigma_1, \ldots \sigma_{j+1}, S^{OPT}) \geq \frac{t}{(1+\varepsilon)}$$

$\Downarrow$ Together with $(* *)$

$$\Delta(\sigma_1 \ldots \sigma_{j+1}, S_{j+1}) \leq (\alpha + O(\varepsilon))(1+\varepsilon)$$

output summary

$$= \alpha + O(\varepsilon)$$

✓