# Lecture 6,7:

# Finding approximate median and clustering in sublinear time

Course: Algorithms for Big Data

Instructor: Hossein Jowhari

Department of Computer Science and Statistics
Faculty of Mathematics
K. N. Toosi University of Technology

Spring 2021

# Recap of last lecture

- Finding an approximate median in sublinear time

- $k$-median clustering in sublinear time

# Approximate median

Input: A large set of elements $A = \{a_1, \ldots, a_n\}$. We assume $D$ has a total ordering.

Rank of an element: $rank(x) = |\{y \in A \mid y \leq x\}|$

Median: $\text{med}(A) = x$ where $rank(x) = \lceil \frac{n}{2} \rceil$.

Approximate Median: An $\epsilon$-approximate median of $A$ is an $y \in A$ where

$$\lceil \frac{n}{2} \rceil - \epsilon n \leq rank(y) \leq \lceil \frac{n}{2} \rceil + \epsilon n$$

$$Sorted(A) = b_1, b_2, \ldots, \underbrace{b_{\lceil \frac{n}{2} \rceil - \epsilon n}, \ldots, \overbrace{b_{\lceil \frac{n}{2} \rceil}}^{\text{median}}, \ldots, b_{\lceil \frac{n}{2} \rceil + \epsilon n}}_{\epsilon\text{-approximate medians}}, \ldots, b_{n-1}, b_n$$

# Finding an approximate median via sampling

Algorithm: Sample $s$ elements from $A$ (with replacement) and return the median of the sample set.

Lemma: If $s \geq \frac{7}{\epsilon^2} \ln(\frac{2}{\delta})$, the algorithm returns an $\epsilon$-approximate median with probability at least $1 - \delta$.

Proof: Partition $A$ into 3 groups:

$$A_L = \{x \in A : rank(x) < \lceil \frac{n}{2} \rceil - \epsilon n\}$$

$$A_M = \{x \in A : \lceil \frac{n}{2} \rceil - \epsilon n \leq rank(x) \leq \lceil \frac{n}{2} \rceil + \epsilon n\}$$

$$A_H = \{x \in A : rank(x) > \lceil \frac{n}{2} \rceil + \epsilon n\}$$

Observation: If less than $\frac{s}{2}$ elements from both $A_L$ and $A_H$ are present in the sample set then the median of the sample is an $\epsilon$-approximate median.

Proof: The argument is similar to what we discussed in Lecture 4 (see page 6).

Let $X_i = 1$ if the $i$-th sample is from $A_L$, otherwise $X_i = 0$. $X = \sum_{i=1}^{s} X_i$.

$$E[X] \leq (\frac{1}{2} - \epsilon)s$$

Assume $\epsilon \leq 0.1$. By Chernoff bound,

$$Pr\Big(X \geq \frac{s}{2}\Big) \leq Pr\Big(X \geq (1+\epsilon)E[X]\Big) \leq e^{-\frac{\epsilon^2}{3}(\frac{1}{2}-\epsilon)s} \leq \frac{\delta}{2}$$

By similar argument, if we set $s \geq 7\epsilon^{-2}\ln(\frac{2}{\delta})$ (assuming $\epsilon \leq 0.1$) the probability that the number of elements from $A_H$ in the sample set is at least $\frac{s}{2}$ is bounded by $\delta/2$.

By union bound, number of elements from both $A_L$ and $A_H$ in the sample set is less than $\frac{s}{2}$ with probability at least $1-\delta$.

Therefore with probability $1-\delta$, the output of the algorithm is an $\epsilon$-approximate median of $A$.

Sample complexity: $O(\frac{1}{\epsilon^2}\ln(\frac{1}{\delta}))$

Homework: Generalize this result to the problem of finding an element with (approximate) rank $t$.

# $k$-median clustering

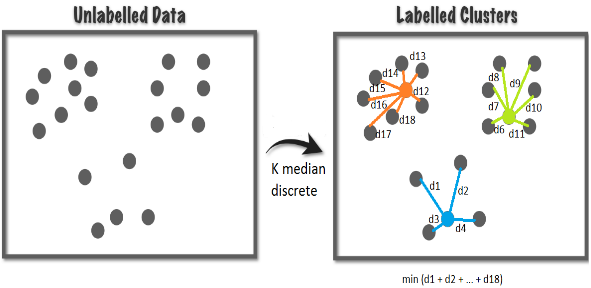
$k$-median clustering problem: Given a metric $(X, d)$ where $X$ is a finite set of data points and $d$ is a distance defined over $X$, in the (discrete) $k$-median problem, the goal is to select $k$ center points $c_1, \ldots, c_k$ from $X$, so that the sum of distances to the closest center is minimized.

$$X = \{x_1, \ldots, x_n\}$$

$$\min_{c_1,\ldots,c_k \subseteq X} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \{d(x_i, c_j)\}$$

Note: In a metric space, the distance is a symmetric function and the triangle inequality holds.

Note: If $|X| = n$, the metric $(X, d)$ can be represented by a symmetric $n$ by $n$ matrix.

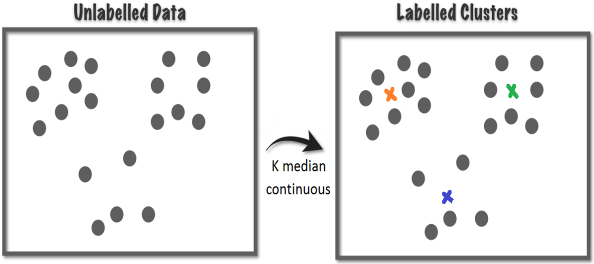



Note: The problem is equivalent to the problem of minimizing the average distance to the closest center.

$$\min_{c_1,\ldots,c_k \subseteq X} \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \{d(x_i, c_j)\}$$

# Continuous $k$-median problem

In the continuous version, the finite set of points $X$ lie in a continuous space (for example $X \subset \mathbb{R}^d$ with the Euclidean distance.) Here we are allowed to choose the $k$ centers from the entire space, not just from the given points $X$.




Note: Both discrete and continuous versions of $k$-median clustering are NP-hard problems. It means, assuming $NP \neq P$, there is no polynomial time algorithm for finding an optimal $k$-median clustering.

# Some algorithmic facts

- Trivially, there is a $O(kn^{k+1})$ time algorithm for finding an optimal $k$-median clustering (discrete version). why?

  There are $\binom{n}{k} = O(n^k)$ ways for selecting the centers.

- The problem is NP-hard even for points in $\mathbb{R}^2$.

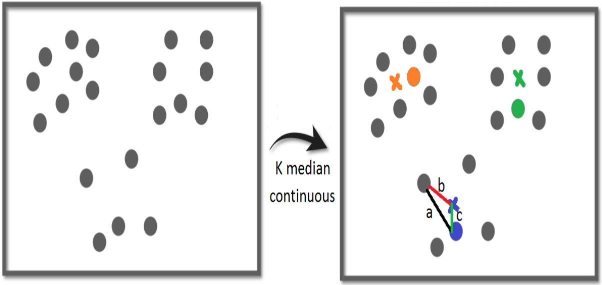
- There is a polynomial time approximation algorithm for $k$-median clustering that returns a solution with cost at most $\alpha = 2.611$ times the optimal cost.

- There is $O(n \log n \log k)$ time constant factor approximation algorithm for $k$-median clustering when the points lie in $\mathbb{R}^d$ with constant $d$.

Lemma: An optimal solution for the discrete version is a $2$-factor approximation solution for the continuous version.

Proof: Use triangle inequality.

Replace each optimal continuous center with its closest point in $X$. See the figure below.

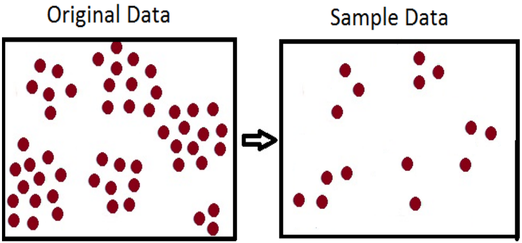



$$a \leq b + c, \quad c \leq b$$

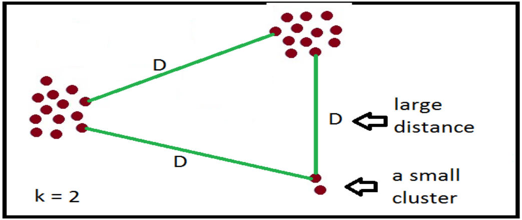
$$\Rightarrow \quad a \leq 2b$$

Corollary: Any $\alpha$-factor approximation algorithm for the discrete version is a $2\alpha$-factor approximation algorithm for the continuous version.

# Sublinear time clustering via sampling




Is the sample a good representative of the whole data?




In general, we need to see the whole data to get a good approximation.

If we make certain assumptions about the data, we may hope that a small sample is a good representative of the whole.

Some algorithmic results in this direction:

- There is a polynomial-time randomized algorithm with query complexity $\tilde{O}(\frac{D^2}{\epsilon^2} k \ln(\frac{n}{\delta}))$ that returns a solution with cost at most $O(OPT) + \epsilon n$ with probability $1 - \delta$. Here $D$ is the diameter of the points. Mishra, Oblinger, Pitt, 2001.

- There is a $O(\frac{k^3}{\epsilon^2} \log^3 k)$ time randomized algorithm that returns a solution with cost $O(OPT)$ under the assumption that every optimal cluster is of size at least $\Omega(\frac{n\epsilon}{k})$. Meyerson et al.2004

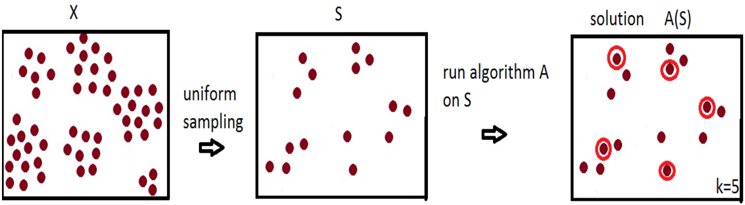
# Mishra, Oblinger, Pitt (MOP)'s Algorithm

Assumption: Suppose there is a deterministic $\alpha$-factor approximation algorithm $A$ for the $k$-median clustering problem that runs in $T(n, k, \alpha)$ time.

MOP's Idea:

- Fix $\epsilon \in (0,1)$ and $\delta \in (0,1)$.
- Pick a sample $S$ of size $s \geq \frac{(\alpha D)^2}{\epsilon^2} k \ln(\frac{n}{\delta})$ from the points $X$. Here $D$ is the diameter of the input points.
- Run algorithm $A$ on the sample $S$ and return the solution.

Let $A(S)$ be the solution ($k$ centers) reported by the approximation algorithm $A$ on the sample set $S$.




$$cost(OPT(X)^{\text{avg}}) = \frac{1}{n} \min_{c_1,\ldots,c_k \subseteq X} \sum_{x \in X} \min_{j=1,\ldots,k} \{d(x, c_j)\}$$

$$cost(A(S)^{\text{avg}}) = \tfrac{1}{s} \textstyle\sum_{x \in S} \min_{c_j \in A(S)} \{d(x, c_j)\}$$

Claim: With probability at least $1 - \delta$, we have

$$cost(A(S)^{\text{avg}}) \leq 2\alpha \, cost(OPT(X)^{\text{avg}}) + \epsilon$$

Fact: (Haussler/Pollard) Let $F$ be a finite set of functions on $X$ with $0 \leq f(x) \leq M$ for all $f \in F$ and $x \in X$. Let $x_1, \ldots, x_m$ be a sequence of $m$ samples drawn independently and identically from $X$ and let $\epsilon > 0$. Let

$$E_T(f) = \frac{1}{|T|} \sum_{x \in T} f(x) \qquad \text{(the average of } f \text{ on } T\text{)}$$

If $m \geq \frac{M^2}{2\epsilon^2} \ln(\frac{2|F|}{\delta})$ then

$$Pr\Big(\exists f \in F \text{ where } |E_X(f) - E_S(f)| \geq \epsilon\Big) \leq \delta.$$

Proof: Use additive Chernoff bound (Homework).

Observation: Every choice of $k$ centers $\{c_1, \ldots, c_k\} \subseteq X$ defines a function

$$f_{c_1,\ldots,c_k}(x) = \min_{j=1,\ldots,k} \{d(x, c_j)\}$$

$$cost_X(\{c_1, \ldots, c_k\}^{\text{avg}}) = \frac{1}{|X|} \sum_{x \in X} f_{c_1,\ldots,c_k}(x) = E_X(f_{c_1,\ldots,c_k})$$

Observation: Let

$$M = \max_{\{c_1,\ldots,c_k\} \subseteq X, x \in X} \{f_{c_1,\ldots,c_k}(x)\}.$$

We have $M \leq D$ where $D$ is the diameter of $X$ (largest distance in $X$.)

According to Haussler/Pollard if we set the number of samples $s \geq \frac{D^2}{2\epsilon^2} \ln(\frac{2|F|}{\delta})$ where

$$F = \{f_{c_1,\ldots,c_k} \mid \{c_1, \ldots, c_k\} \subseteq X\}, \qquad |F| = \binom{n}{k}$$

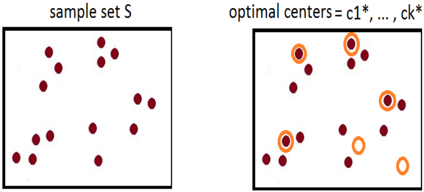
then with probability $1 - \delta$ for all $f_{c_1,\ldots,c_k} \in F$ we have

$$|E_X(f_{c_1,\ldots,c_k}) - E_S(f_{c_1,\ldots,c_k})| \leq \epsilon.$$

In other words, if we replace $X$ with the sample set $S$, the (average) cost of any clustering on $S$ (using a set of centers) will be close to the corresponding (average) cost on $X$. In particular it means good centers for $S$ will be good centers for $X$ (with some additive error.)

Assumption: In the rest of the analysis, we assume the good event happens and all functions in $F$ have near equal values on $X$ and $S$.

Let $c_1^*, \ldots, c_k^*$ be the optimal centers for $X$.
Note, some of the centers in $\{c_1^*, \ldots, c_k^*\}$ might not belong to $S$.




Let $z_1^*, \ldots, z_k^*$ be the optimal centers for the continuous $k$-median on $S$.

Let $z_1, \ldots, z_k$ be the optimal centers for the discrete $k$-median on $S$.

Let $a_1, \ldots, a_k$ be the centers found by the $\alpha$-approximation algorithm on $S$.

We have the following observations:

- $\forall f_{c_1,\dots,c_k} \in F,\ |E_X(f_{c_1,\dots,c_k}) - E_S(f_{c_1,\dots,c_k})| \le \epsilon.$ (1)

- $\underbrace{E_S(f_{z_1^*,\dots,z_k^*})}_{\text{continuous cost}} \le \underbrace{E_S(f_{z_1,\dots,z_k})}_{\text{discrete cost}} \le 2\,\underbrace{E_S(f_{z_1^*,\dots,z_k^*})}_{\text{continuous cost}}$ (2)

- $\underbrace{E_S(f_{z_1,\dots,z_k})}_{\text{discrete cost}} \le \underbrace{E_S(f_{a_1,\dots,a_k})}_{\text{algorithm cost}} \le \alpha\,\underbrace{E_S(f_{z_1,\dots,z_k})}_{\text{discrete cost}}$ (3)

- $(2), (3) \Rightarrow$

$$\underbrace{E_S(f_{z_1^*,\dots,z_k^*})}_{\text{continuous cost}} \le \underbrace{E_S(f_{a_1,\dots,a_k})}_{\text{algorithm cost}} \le 2\alpha\,\underbrace{E_S(f_{z_1^*,\dots,z_k^*})}_{\text{continuous cost}} \quad (4)$$

- $\underbrace{E_S(f_{z_1^*,\dots,z_k^*})}_{\text{continuous cost}} \le E_S(f_{c_1^*,\dots,c_k^*})$ (5)

- $(1), (4), (5) \Rightarrow$

$$E_X(f_{a_1,\dots,a_k}) - \epsilon \le \underbrace{E_S(f_{a_1,\dots,a_k})}_{\text{algorithm cost}} \le 2\alpha E_S(f_{c_1^*,\dots,c_k^*}) \tag{6}$$

- Since $c_1^*, \dots, c_k^*$ are the optimal centers for $X$,

$$E_X(f_{c_1^*,\dots,c_k^*}) - \epsilon \le \underbrace{E_S(f_{a_1,\dots,a_k})}_{\text{algorithm cost}} \le 2\alpha E_S(f_{c_1^*,\dots,c_k^*}) \tag{7}$$

- $(1) \Rightarrow$

$$E_X(f_{c_1^*,\dots,c_k^*}) - \epsilon \le \underbrace{E_S(f_{a_1,\dots,a_k})}_{\text{algorithm cost}} \le 2\alpha(E_X(f_{c_1^*,\dots,c_k^*}) + \epsilon)$$

- $E_X(f_{c_1^*,\dots,c_k^*}) - \epsilon \le \underbrace{E_S(f_{a_1,\dots,a_k})}_{\text{algorithm cost}} \le 2\alpha E_X(f_{c_1^*,\dots,c_k^*}) + 2\alpha\epsilon$

Since $|F| \leq n^k$, if we replace $\epsilon$ by $\frac{\epsilon}{2\alpha}$ and choose $s \geq \frac{2\alpha^2 D^2}{\epsilon^2} k \ln(\frac{2n}{\delta})$, with probability at least $1 - \delta$, we get

$$\underbrace{E_X(f_{c_1^*,\ldots,c_k^*})}_{\text{optimal cost on } X} - \frac{\epsilon}{2\alpha} \leq \underbrace{E_S(f_{a_1,\ldots,a_k})}_{\text{algorithm cost}} \leq 2\alpha \underbrace{E_X(f_{c_1^*,\ldots,c_k^*})}_{\text{optimal cost on } X} + \epsilon$$

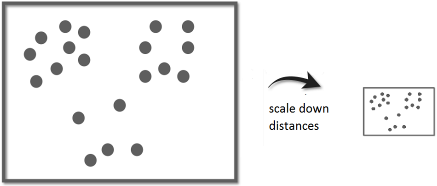
$$\Downarrow$$

$$\underbrace{E_X(f_{c_1^*,\ldots,c_k^*})}_{\text{optimal cost on } X} - \epsilon \leq \underbrace{E_S(f_{a_1,\ldots,a_k})}_{\text{algorithm cost}} \leq 2\alpha \underbrace{E_X(f_{c_1^*,\ldots,c_k^*})}_{\text{optimal cost on } X} + \epsilon$$

# Few Remarks and Questions

- The running time of the final algorithm depends on sample size $s$ and the running time of the $\alpha$-factor approximation algorithm $A$.

- Why don't we scale down all distances so that the diameter of the points is reduced to $1$? At first glance, this seems to bring down the sampling complexity to $O(\frac{\alpha^2 k \ln(\frac{n}{\delta})}{\epsilon^2})$. Why do you think this idea fails?



# Better analysis by Czumaj and Sohler

Czumaj and Sohler have shown taking $O(\frac{D}{\epsilon^2}(k + \ln(\frac{1}{\delta})))$ sample points is enough to find a solution with the same quality.

**Sublinear-Time Approximation for Clustering via Random Sampling ⋆**


Artur Czumaj[1] and Christian Sohler[2]

[1] Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA. czumaj@cis.njit.edu

[2] Heinz Nixdorf Institute and Department of Computer Science, University of Paderborn, D-33102 Paderborn, Germany. csohler@uni-paderborn.de



**Abstract.** In this paper we present a novel analysis of a random sampling approach for three clustering problems in metric spaces: *k-median*, *min-sum k-clustering*, and *balanced k-median*. For all these problems we consider the following simple sampling scheme: select a small sample set of points uniformly at random from $V$ and then run some approximation algorithm on this sample set to compute an approximation of the best possible clustering of this set. Our main technical contribution is a significantly strengthened analysis of the approximation guarantee by this scheme for the clustering problems.

# Something to think about

Does the analysis work for other clustering objectives such as $k$-center or $k$-means?

$k$-center clustering:

$$\min_{c_1,\dots,c_k \subseteq X} \ \max_{x \in X} \ \min_{j=1,\dots,k} \{d(x, c_j)\}$$

$k$-means clustering:

$$\min_{c_1,\dots,c_k \subseteq \mathbb{R}^d} \ \sum_{x \in X \subseteq \mathbb{R}^d} \ \min_{j=1,\dots,k} \{\|x - c_j\|^2\}$$