Abstract

In this paper, one-dimensional $k$-medians clustering problem is considered in the context of zero-sum game between players choosing a sample and partitioning it into clusters, respectively. For any sample size $n$ and $k > 1$, an attainable guaranteed value of the clustering accuracy $0.5n/(2k - 1)$ (the low value of an appropriate game) is provided for samples taken from the segment $[0,1]$.

1 Introduction

In data analysis, $k$-medians clustering problem is regarded as one of the famous center-based metric clustering problems, whose instance can be defined as follows. For a given number $k \geq 1$ and a finite sample $\xi = (x_1, \ldots, x_n)$ taken from a metric space $(X, \rho)$, it is required to find a partition of $\mathbb{N}_n = \{1, \ldots, n\}$ onto $k$ clusters $C_1, \ldots, C_k$ and, for any $j$-th cluster, to point out an appropriate center $c_j$ such that

$$\sum_{j=1}^k \sum_{i \in C_j} \rho(x_i, c_j) = \sum_{i=1}^n \min\{\rho(x_i, c_1), \ldots, \rho(x_i, c_k)\} \to \min.$$  

Equation (1) evidently implies that, for any $j$, the point $c_j \in \text{Arg min}\left\{\sum_{i \in C_j} \rho(x_i, c) : c \in X\right\}$, i.e. $c_j$ is a median of the subsample $\xi_j = (x_i : i \in C_j)$.

As a combinatorial optimization problem, $k$-medians is shown to be intractable [Guruswami and Indyk, 2003] even for the Euclidean metric and has no PTAS, unless $P = NP$. For $d$-dimensional Euclidean spaces there

---

1 If $k$ is a part of an instance.
are known numerous approximation results. For instance, in [Kumar et al., 2010], for any fixed \( k \), randomized LTAS with time complexity of \( O(2^{(k/\varepsilon)^{O(1)}} \cdot dn) \) is proposed. On the basis of the famous coresets technique, in [Har-Peled and Mazumdar, 2004], RPTAS with polynomially depending on the number of clusters \( k \) time complexity bound \( O(n + \rho(k \log n)^{O(1)}) \), where \( \rho = \exp(O(1 - \log \varepsilon)/\varepsilon^{d-1}) \) is proposed. For \( d = 1 \), \( k \)-medians problem is polynomially (and very efficiently) solvable. To date, the most efficient exact algorithm with time complexity \( O(n \log n + kn) \) is proposed in [Grønlund et al., 2017].

Among others, the setting, where it is required to obtain a guaranteed accuracy of clustering for a fixed number of clusters \( k \) and an arbitrary sample, is valuable ([Ben-David, 2015, Khachai and Neznakhina, 2017]) for applications in combinatorial optimization and data analysis. In this paper, we study such a setting for the 1d-case of the \( k \)-medians clustering problem.

2 Problem Statement and the Main Result

We consider the following two-player zero-sum game induced by \( k \)-medians clustering. There are two players placing points in the unit segment of the real line. Strategies of the first player are samples \( \xi = (x_1, \ldots, x_n), x_i \in [0,1] \) of some given size \( n \). Strategies of the second one are \( k \)-tuples \( \sigma = (c_1, \ldots, c_k), c_i \in [0,1] \). The payoff function \( F(\xi, \sigma) = \sum_{i=1}^{n} \min\{|x_i - c_1|, \ldots, |x_i - c_k|\} \). Goals of the first and the second players are to find the lower and the higher values of the game, respectively.

It is easy to verify that, for any \( k > 1 \) and \( n > 0 \), the game has no value, i.e. \( v_*(n, k) < v^*(n, k) \). For many reasons arising from applications in data analysis, combinatorial optimization, and computational geometry, it is important to have an upper bound for \( v_*(n) \), which means the guaranteed accuracy of \( k \)-medians clustering of an appropriate \( n \)-points sample. Although, \( v^*(n, k) \) can obviously be taken as an upper bound, for large values of \( n \) it is imprecise and should be replaced with more accurate one.

In this paper, we propose an attainable upper bound \( B(n, k) \) for \( v_*(n, k) \). Actually, to any \( n > 0, k > 1 \), and \( \xi \in [0,1]^n \), we show how to assign an appropriate \( k \)-tuple \( \sigma_\xi = (c_1, \ldots, c_k) \), i.e. how to construct a clustering \( C_1, \ldots, C_k \) with medians \( c_1, \ldots, c_k \), such that

\[
\inf_{\sigma \in [0,1]^k} F(\xi, \sigma) \leq F(\xi, \sigma_\xi) \leq B(n, k).
\]

Theorem.

(i) For any \( k > 1 \), \( n > 0 \), and sample \( \xi = (x_1, \ldots, x_n), x_i \in [0,1], i \in \mathbb{N}_n \), there exists the \( k \)-tuple \( \sigma_\xi = (c_1, \ldots, c_k), c_j \in [0,1], j \in \mathbb{N}_k \), such that

\[
F(\xi, \sigma_\xi) \leq \frac{n}{2(2k - 1)}.
\]  

(ii) For any \( k > 1 \), there is \( \bar{n} = \bar{n}(k) \) such that, for all \( n > \bar{n} \), bound (2) is attained at some sample \( \xi = \xi(k, n) \).

Postponing the rigorous proof to the forthcoming paper, we restrict ourselves to some suggestive thoughts. To put it simple, we consider the case of \( k = 2 \).

3 Proof Sketch for \( k = 2 \)

We start with the following simple upper bound

3.1 Naïve Upper Bound

It can be assumed that the second player always adheres to the following strategy. He splits the segment \([0,1]\) onto two equal parts and put \( c_1 \) and \( c_2 \) at the centers of each part as it is shown in Fig. 1.

Obviously, in this case, for any \( x \in [0,1] \), \( \min\{|x - c_1|, |x - c_2|\} \leq 1/4 \). Therefore, regardless of the choice \( \xi = (x_1, \ldots, x_n) \) of the first player, \( \sum_{i=1}^{n} \min\{|x_i - c_1|, |x_i - c_2|\} \leq n/4 \), i.e. \( B(n, 2) \leq n/4 \). Since, to complete the first point of the proof (for the considered case \( k = 2 \)), we need to show that \( B(n, 2) \leq n/6 \), we need further improvements.
ultimately and obeys the equation

Then, for the median \( m \) which is equivalent to \( m \) constraints) in problem (4) by half. Indeed, suppose, \( v \)

Thus, \( v_n(2) = \sup_{\xi \in [0,1]^n} \Phi(\xi) \) is an optimum value of linear program (4)

Thus, \( v_n(2) = \max u \)

\[
\Phi(\xi) = \min \left\{ \sum_{i \in C_1} |x_i - c_1| + \sum_{i \in C_2} |x_i - c_2| : C_1 \cup C_2 = \mathbb{N}_n \right\} = \min \left\{ - \sum_{i=1}^{m_1/2} x_i + \sum_{i=m_1/2+1}^{m_1} x_i - \sum_{i=1}^{m_2/2} x_{i+m_1} + \sum_{i=m_2/2+1}^{m_2} x_{i+m_1} : m_1 + m_2 = n \right\}.
\]

Therefore, for a given sample \( \xi \), \( \Phi(\xi) = \inf_{\sigma = (c_1, c_2)} F(\xi, \sigma) \) depends on choice of partitions \( C_1 \cup C_2 = \mathbb{N}_n \) ultimately and obeys the equation

Thus, \( v_n(2) = \sup_{\xi \in [0,1]^n} \Phi(\xi) \) is an optimum value of linear program (4)

Further, guided by the symmetry argument, we can reduce the number of variables (and also, the number of constraints) in problem (4) by half. Indeed, suppose, \( \xi' = (x_1', \ldots, x_n') \) is an optimal solution of (4). Then, by symmetry, \( \xi'' = (1 - x_n', \ldots, 1 - x_1') \) is an optimal solution of (4) as well. Convexity of the optimal set of (4) implies that \( \xi = (\xi' + \xi'')/2 \), each whose entry is defined by the formula \( x_i = (1 + x_i' - x_{n+1-i}')/2 \) is also an optimal solution. Since \( x_i + x_{n+1-i} = 1 \), hereinafter, we reduce the number of variables to \( n/2 \). Moreover, for odd \( n \), \( x_{n/2} = 1/2 \).

To show that \( B(n, 2) \leq n/6 \), we study all cases for \( (n \mod 6) \).

Case \( n = 6t \):

Consider the constraint of (4) defined by \( m_1 = 2t \) and \( m_2 = 4t \).

which is equivalent to \( u + 2 \sum_{i=1}^{t} x_i \leq t \). Since all \( x_i \geq 0, u \leq t = n/6 \), and we are done.

Case \( n = 6t + 1 \):

Here, we consider two constraints of (4), defined by \( m_1 = 2t, m_2 = 4t + 1 \) and \( m_1 = 2t + 1, m_2 = 4t \), respectively. They are

\[
- \sum_{i=1}^{t} x_i + \sum_{i=t+1}^{2t} x_i - \sum_{i=2t+1}^{3t} x_i - \sum_{i=2t+1}^{4t} (1-x_i) + \sum_{i=1}^{2t} (1-x_i) \geq u
\]

Figure 1: Simple upper bound

3.2 Reducing to Linear Program

Hereinafter, without loss of generality, we assume that any sample \( \xi = (x_1, \ldots, x_n) \) contains points \( x_i \) in ascending order. Moreover, we assume that any cluster \( C = \{i_1, \ldots, i_m\} \subset \mathbb{N}_n \) inherits this property, i.e. \( x_{i_1} \leq \ldots \leq x_{i_m} \).

Then, for the median \( c \) of the cluster \( C \) we have

Therefore, for a given sample \( \xi \), \( \Phi(\xi) = \inf_{\sigma = (c_1, c_2)} F(\xi, \sigma) \) depends on choice of partitions \( C_1 \cup C_2 = \mathbb{N}_n \) ultimately and obeys the equation

Thus, \( v_n(2) = \sup_{\xi \in [0,1]^n} \Phi(\xi) \) is an optimum value of linear program (4)

Further, guided by the symmetry argument, we can reduce the number of variables (and also, the number of constraints) in problem (4) by half. Indeed, suppose, \( \xi' = (x_1', \ldots, x_n') \) is an optimal solution of (4). Then, by symmetry, \( \xi'' = (1 - x_n', \ldots, 1 - x_1') \) is an optimal solution of (4) as well. Convexity of the optimal set\(^2\) of (4) implies that \( \xi = (\xi' + \xi'')/2 \), each whose entry is defined by the formula \( x_i = (1 + x_i' - x_{n+1-i}')/2 \) is also an optimal solution. Since \( x_i + x_{n+1-i} = 1 \), hereinafter, we reduce the number of variables to \( n/2 \). Moreover, for odd \( n \), \( x_{n/2} = 1/2 \).

To show that \( B(n, 2) \leq n/6 \), we study all cases for \( (n \mod 6) \).

Case \( n = 6t \):

Consider the constraint of (4) defined by \( m_1 = 2t \) and \( m_2 = 4t \).

which is equivalent to \( u + 2 \sum_{i=1}^{t} x_i \leq t \). Since all \( x_i \geq 0, u \leq t = n/6 \), and we are done.

Case \( n = 6t + 1 \):

Here, we consider two constraints of (4), defined by \( m_1 = 2t, m_2 = 4t + 1 \) and \( m_1 = 2t + 1, m_2 = 4t \), respectively. They are

\[
- \sum_{i=1}^{t} x_i + \sum_{i=t+1}^{2t} x_i - \sum_{i=2t+1}^{3t} x_i - \sum_{i=2t+1}^{4t} (1-x_i) + \sum_{i=1}^{2t} (1-x_i) \geq u
\]

\(^2\)The set of optimal solutions
and

\[- \sum_{i=1}^{t} x_i + \sum_{i=t+2}^{2t+1} x_i - \sum_{i=2t+2}^{2t+3} x_i - \frac{1}{2} \sum_{i=2t+1}^{2t+3} (1 - x_i) + \sum_{i=1}^{2t+1} (1 - x_i) \geq u.\]

After the equivalent transformation, we obtain the subsystem

\[
\begin{align*}
    & u + 2 \sum_{i=1}^{t} x_i + x_{2t+1} \leq t + \frac{1}{2} \\
    & u + 2 \sum_{i=1}^{t} x_i + x_{t+1} - 2x_{2t+1} \leq t - \frac{1}{2},
\end{align*}
\]

which implies

\[3u + 6 \sum_{i=1}^{t} x_i + x_{t+1} \leq 3t + 1/2 \quad \text{and} \quad u \leq t + 1/6 = n/6.\]

In case \(n = 6t + 2\)

we take constraints defined by \(m_1 = 2t + 1, m_2 = 4t + 1\) and \(m_1 = 2t, m_2 = 4t + 2\):

\[
\begin{align*}
    & - \sum_{i=1}^{t} x_i + \sum_{i=t+2}^{2t+1} x_i - \sum_{i=2t+2}^{2t+3} x_i - \frac{1}{2} \sum_{i=2t+1}^{2t+3} (1 - x_i) + \sum_{i=1}^{2t+1} (1 - x_i) \geq u \\
    & - \sum_{i=1}^{t} x_i + \sum_{i=t+1}^{2t+1} x_i - \sum_{i=2t+2}^{2t+3} x_i - \frac{1}{2} \sum_{i=2t+1}^{2t+3} (1 - x_i) + \sum_{i=1}^{2t+1} (1 - x_i) \geq u.
\end{align*}
\]

Transformed

\[
\begin{align*}
    & u + 2 \sum_{i=1}^{t} x_i - x_{2t+1} \leq t \\
    & u + 2 \sum_{i=1}^{t} x_i + 2x_{2t+1} \leq t + 1,
\end{align*}
\]

they imply

\[3u + 6 \sum_{i=1}^{t} x_i \leq 3t + 1 \quad \text{i.e.} \quad u \leq t + 1/3 = n/6.\]

Case \(n = 6t + 3\)

is similar to the case \(n = 6t\). Here, to obtain the desired bound, it is enough to consider the single constraint defined by \(m_1 = 2t + 1\) and \(m_2 = 4t + 2\)

\[
- \sum_{i=1}^{t} x_i + \sum_{i=t+2}^{2t+1} x_i - \sum_{i=2t+2}^{2t+3} x_i - \frac{1}{2} \sum_{i=2t+1}^{2t+3} (1 - x_i) + \sum_{i=1}^{2t+1} (1 - x_i) \geq u. \quad (5)
\]

Being transformed, (5) becomes

\[u + 2 \sum_{i=1}^{t} x_i + x_{t+1} \leq t + 1/2,\]

which implies \(u \leq t + 1/2 = n/6.\)

In case \(n = 6t + 4\)

we convolve again two appropriate constraints defined by \(m_1 = 2t + 1, m_2 = 4t + 3\) and \(m_1 = 2t + 2, m_2 = 4t + 2\)

\[
\begin{align*}
    & - \sum_{i=1}^{t} x_i + \sum_{i=t+2}^{2t+1} x_i - \sum_{i=2t+2}^{2t+3} x_i - \frac{1}{2} \sum_{i=2t+1}^{2t+3} (1 - x_i) + \sum_{i=1}^{2t+1} (1 - x_i) \geq u \\
    & - \sum_{i=1}^{t} x_i + \sum_{i=t+2}^{2t+2} x_i - \sum_{i=2t+3}^{2t+4} x_i - \frac{1}{2} \sum_{i=2t+2}^{2t+4} (1 - x_i) + \sum_{i=1}^{2t+2} (1 - x_i) \geq u,
\end{align*}
\]

325
which, after the equivalent transformation give the subsystem
\[
\begin{cases}
u + 2 \sum_{i=1}^{t} x_i + x_{2t+2} \leq t + 1 \\
u + 2 \sum_{i=1}^{t+1} x_i - x_{2t+2} \leq t
\end{cases}
\]
implying
\[
3u + 6 \sum_{i=1}^{t} x_i + 2x_{t+1} \leq 3t + 2 \quad \text{i.e.} \quad u \leq t + 2/3 = n/6.
\]
Finally, in case \(n = 6t + 5\)
transforming the constraints defined by \(m_1 = 2t + 2, m_2 = 4t + 3\) and \(m_1 = 2t + 1, m_2 = 4t + 4\)
\[
\begin{align*}
- \sum_{i=1}^{t+1} x_i + \sum_{i=t+2}^{2t+2} x_i - \sum_{i=2t+3}^{3t+2} x_i - \frac{1}{2} \sum_{i=2t+3}^{3t+2} (1 - x_i) + \sum_{i=1}^{2t+1} (1 - x_i) \geq u \\
- \sum_{i=1}^{t+1} x_i - \sum_{i=t+2}^{2t+1} x_i - \sum_{i=2t+2}^{3t+2} x_i - \frac{1}{2} \sum_{i=2t+3}^{3t+2} (1 - x_i) + \sum_{i=1}^{2t+2} (1 - x_i) \geq u
\end{align*}
\]
we obtain the subsystem
\[
\begin{cases}
u + 2 \sum_{i=1}^{t} x_i - x_{2t+2} \leq t + \frac{1}{2} \\
u + 2 \sum_{i=1}^{t} x_i + x_{t+1} + x_{2t+2} \leq t + \frac{3}{2},
\end{cases}
\]
which, being convolved, gives us
\[
3u + 6 \sum_{i=1}^{t} x_i + 5x_{t+1} \leq 3t + 5/2 \quad \Rightarrow \quad u \leq t + 5/6 = n/6.
\]
Thus, we completely proved point (i) of Theorem for the case of \(k = 2\).

### 3.3 Attainability

Now, we show that for any \(n \geq 12\) inequality (2) is tight. Consider the following configuration given by locations \(p_1, \ldots, p_5\)

\[
\begin{array}{cccccc}
p_1 & p_2 & p_3 & p_4 & p_5 \\
0 & \frac{1}{3} & \frac{1}{2} & \frac{2}{3} & 1
\end{array}
\]

**Figure 2:** The configuration

Place \(n = 4\left\lfloor \frac{n}{4} \right\rfloor + \left\{ \frac{n}{4} \right\}\) points at the locations \(p_1, \ldots, p_5\) with multiplicities presented at Fig. 3

\[
\begin{array}{cccccc}
p_1 & p_2 & p_3 & p_4 & p_5 \\
\left\lfloor \frac{n}{4} \right\rfloor & \left\lfloor \frac{n}{4} \right\rfloor & \left\{ \frac{n}{4} \right\} & \left\lfloor \frac{n}{4} \right\rfloor & \left\lfloor \frac{n}{4} \right\rfloor
\end{array}
\]

**Figure 3:** Placing the points

Since \(n \geq 12\), the multiplicities of points located at \(p_1, p_2, p_4\), and \(p_5\) are at least 3 and at most 3 points are located at \(p_3\). By the symmetry of the sample obtained, there are two best options to partition it into two clusters \(C_1 = \{1, \ldots, \lfloor n/4 \rfloor\}, C_2 = \{\lfloor n/4 \rfloor + 1, \ldots, n\}\) and \(C_1 = \{1, \ldots, 2\lfloor n/4 \rfloor\}, C_2 = \{2\lfloor n/4 \rfloor + 1, \ldots, n\}\) (see Fig.4).
Figure 4: Two ways of possible clustering

Let us calculate the cost $F(\xi, \sigma)$ for each option. In the first case

$$F(\xi, \sigma) = \sum_{i \in C_2} |x_i - c_2|,$$

where $c_2 = p_4$ (since $n > 12$). Therefore,

$$F(\xi, \sigma) = \left\lfloor \frac{n}{4} \right\rfloor \frac{1}{3} + \frac{n}{6} + \left\lfloor \frac{n}{4} \right\rfloor \frac{1}{3} + \left\lfloor \frac{n}{4} \right\rfloor \frac{1}{3} = \frac{4}{6} \left\lfloor \frac{n}{4} \right\rfloor + \frac{n}{6} = \frac{n}{6}.$$

Consider the second case. Here, again $c_2 = p_4$. Therefore,

$$F(\xi, \sigma) = \left\lfloor \frac{n}{4} \right\rfloor \frac{1}{3} + \frac{n}{6} + \left\lfloor \frac{n}{4} \right\rfloor \frac{1}{3} = \frac{n}{6},$$

i.e. Theorem is completely proved so as point (ii).

Acknowledgements

This research was supported by Russian Foundation for Basic Research, projects no. 16-07-00266 and no. 17-08-01385.

References


