Attainable Best Guarantee for the Accuracy of k-medians Clustering in $[0,1]$

Michael Khachay
Krasovsky Institute of Mathematics and Mechanics,
Ural Federal University, Ekaterinburg, Russia
Omsk State Technical University, Omsk, Russia
mkhachay@imm.uran.ru

Vasiliy Pankratov
Krasovsky Institute of Mathematics and Mechanics,
pankratov.vs@gmail.com

Daniel Khachay
Krasovsky Institute of Mathematics and Mechanics,
Ural Federal University, Ekaterinburg, Russia
dmx@imm.uran.ru

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, ρ), 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

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

Equation (1) evidently implies that, for any j, the point $c_j \in \operatorname{Arg\,min}\left\{\sum_{i \in C_j} \rho(x_i, c) : c \in X\right\}$, 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[1] 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

In: Yu. G. Evtushenko, M. Yu. Khachay, O. V. Khamisov, Yu. A. Kochetov, V.U. Malkova, M.A. Posypkin (eds.): Proceedings of the OPTIMA-2017 Conference, Petrovac, Montenegro, 02-Oct-2017, published at http://ceur-ws.org

[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 $\xi = (x_1, \dots, x_n)$, $x_i \in [0, 1]$ of some given size n. Strategies of the second one are k-tuples $\sigma = (c_1, \dots, c_k)$, $c_i \in [0, 1]$. The payoff function $F(\xi, \sigma) = \sum_{i=1}^{n} \min\{|x_i - c_1|, \dots, |x_i - c_k|\}$. Goals of the first and the second players are to find the lower

$$v_*(n, k) = \sup_{\xi \in [0,1]^n} \inf_{\sigma \in [0,1]^k} F(\xi, \sigma)$$

and the higher

$$v^*(n, k) = \inf_{\sigma \in [0,1]^k} \sup_{\xi \in [0,1]^n} F(\xi, \sigma)$$

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 $\sigma_\xi = (c_1, \dots, c_k)$, i.e. how to construct a clustering $C_1, \dots, C_k$ with medians $c_1, \dots, c_k$, 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 $\xi = (x_1, \dots, x_n)$, $x_i \in [0, 1]$, $i \in \mathbb{N}_n$, there exists the k-tuple $\sigma_\xi = (c_1, \dots, c_k)$, $c_j \in [0, 1]$, $j \in \mathbb{N}_k$, such that*

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

(ii) *For any $k > 1$, there is $\tilde{n} = \tilde{n}(k)$ such that, for all $n > \tilde{n}$, 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 $\xi = (x_1, \dots, x_n)$ 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.

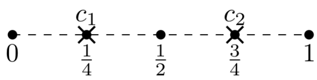

Figure 1: Simple upper bound

3.2 Reducing to Linear Program

Hereinafter, without loss of generality, we assume that any sample $\xi = (x_1, \dots, x_n)$ contains points x_i in ascending order. Moreover, we assume that any cluster $C = \{i_1, \dots, i_m\} \subset \mathbb{N}_n$ inherites this property, i.e. $x_{i_1} \le \dots \le x_{i_m}$. Then, for the median c of the cluster C we have

$$\sum_{l=1}^{m} |x_{i_l} - c| = \sum_{l=1}^{\lfloor m/2 \rfloor} (c - x_{i_l}) \; + \sum_{l=\lceil m/2 \rceil+1}^{m} (x_{i_l} - c) = -\sum_{l=1}^{\lfloor m/2 \rfloor} x_{i_l} \; + \sum_{l=\lceil m/2 \rceil+1}^{m} x_{i_l}. \tag{3}$$

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

$$\Phi(\xi) = \min\left\{ \sum_{i \in C_1} |x_i - c_1| + \sum_{i \in C_2} |x_i - c_2| \colon C_1 \cup C_2 = \mathbb{N}_n \right\}$$
$$= \min\left\{ -\sum_{i=1}^{\lfloor m_1/2 \rfloor} x_i \; + \sum_{i=\lceil m_1/2 \rceil+1}^{m_1} x_i - \sum_{i=1}^{\lfloor m_2/2 \rfloor} x_{i+m_1} \; + \sum_{i=\lceil m_2/2 \rceil+1}^{m_2} x_{i+m_1} \colon m_1 + m_2 = n \right\}.$$

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

$$\begin{aligned} v_*(n,2) = \;& \max u \\ & s.t. \\ & -\sum_{i=1}^{\lfloor m_1/2 \rfloor} x_i \; + \sum_{i=\lceil m_1/2 \rceil+1}^{m_1} x_i - \sum_{i=1}^{\lfloor m_2/2 \rfloor} x_{i+m_1} \; + \sum_{i=\lceil m_2/2 \rceil+1}^{m_2} x_{i+m_1} \ge u, \quad (m_1 + m_2 = n), \\ & 0 \le x_1 \le \dots \le x_n \le 1. \end{aligned} \tag{4}$$

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, $\xi' = (x'_1, \dots, x'_n)$ is an optimal solution of (4). Then, by symmetry, $\xi'' = (1 - x'_n, \dots, 1 - x'_1)$ 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 $\lfloor n/2 \rfloor$. Moreover, for odd n, $x_{\lceil n/2 \rceil} = 1/2$.

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

Case $n = 6t$:

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

$$-\sum_{i=1}^{t} x_i + \sum_{i=t+1}^{2t} x_i - \sum_{i=2t+1}^{3t} x_i - \sum_{i=2t+1}^{3t} (1 - x_i) + \sum_{i=1}^{2t} (1 - x_i) \ge u,$$

which is equivalent to $u + 2\sum_{i=1}^{t} x_i \le t$. Since all $x_i \ge 0$, $u \le t = n/6$, 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

$$-\sum_{i=1}^{t} x_i + \sum_{i=t+1}^{2t} x_i - \sum_{i=2t+1}^{3t} x_i - \frac{1}{2} - \sum_{i=2t+2}^{3t} (1 - x_i) + \sum_{i=1}^{2t} (1 - x_i) \ge u$$

[2] The set of optimal solutions

and

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

After the equivalent transformation, we obtain the subsystem

$$\begin{cases} u + 2\sum\limits_{i=1}^{t} x_i + x_{2t+1} \leq t + \frac{1}{2} \\ u + 2\sum\limits_{i=1}^{t} x_i + x_{t+1} - 2x_{2t+1} \leq t - \frac{1}{2}, \end{cases}$$

which implies

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

In case $n = 6t + 2$

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

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

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

Transformed

$$\begin{cases} u + 2\sum\limits_{i=1}^{t} x_i - x_{2t+1} \leq t \\ u + 2\sum\limits_{i=1}^{t} x_i + 2x_{2t+1} \leq t + 1, \end{cases}$$

they imply

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

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 $m_2 = 4t + 2$

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

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 $m_1 = 2t+2, m_2 = 4t+2$

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

$$-\sum_{i=1}^{t+1} x_i + \sum_{i=l+2}^{2t+2} x_i - \sum_{i=2t+3}^{3t+2} x_i - \sum_{i=2t+2}^{3t+2} (1-x_i) + \sum_{i=1}^{2t+1}(1-x_i) \geq u,$$

which, after the equivalent transformation give the subsystem

$$\begin{cases} u + 2\sum_{i=1}^{t} x_i + x_{2t+2} \le t+1 \\ u + 2\sum_{i=1}^{t+1} x_i - 2x_{2t+2} \le t \end{cases}$$

implying

$$3u + 6\sum_{i=1}^{t} x_i + 2x_{t+1} \le 3t+2 \text{ i.e. } u \le t + 2/3 = n/6.$$

Finally, in case $n = 6t+5$

transforming the constraints defined by $m_1 = 2t+2, m_2 = 4t+3$ and $m_1 = 2t+1, m_2 = 4t+4$

$$-\sum_{i=1}^{t+1} x_i + \sum_{i=t+2}^{2t+2} x_i - \sum_{i=2t+3}^{3t+2} x_i - \frac{1}{2} - \sum_{i=2t+3}^{3t+2} (1-x_i) + \sum_{i=1}^{2t+1} (1-x_i) \ge u$$
$$-\sum_{i=1}^{t} x_i + \sum_{i=t+2}^{2t+1} x_i - \sum_{i=2t+2}^{3t+2} x_i - \frac{1}{2} - \sum_{i=2t+3}^{3t+2} (1-x_i) + \sum_{i=1}^{2t+2} (1-x_i) \ge u$$

we obtain the subsystem

$$\begin{cases} u + 2\sum_{i=1}^{t+1} x_i - x_{2t+2} \le t + \frac{1}{2} \\ u + 2\sum_{i=1}^{t} x_i + x_{t+1} + x_{2t+2} \le t + \frac{3}{2}, \end{cases}$$

which, being convolved, gives us

$$3u + 6\sum_{i=1}^{t} x_i + 5x_{t+1} \le 3t + 5/2 \implies u \le t + 5/6 = n/6.$$

Thus, we completely proved point (i) of Theorem for the case of $k = 2$.

3.3 Attainability

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

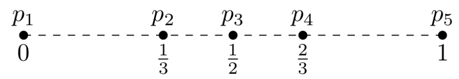

Figure 2: The configuration

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

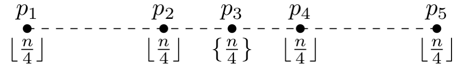

Figure 3: Placing the points

Since $n \ge 12$, 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).

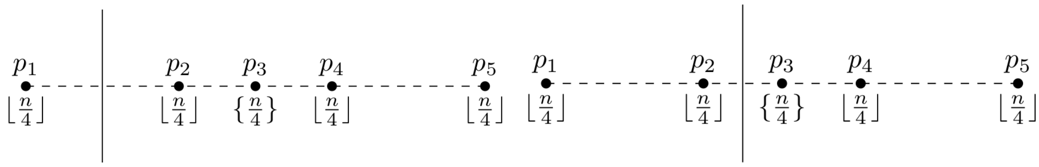

Figure 4: Two ways of possible clustering

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,

$$F(\xi,\sigma) = \left\lfloor\frac{n}{4}\right\rfloor\frac{1}{3} + \left\{\frac{n}{4}\right\}\frac{1}{6} + \left\lfloor\frac{n}{4}\right\rfloor\frac{1}{3} = \left\lfloor\frac{n}{4}\right\rfloor\frac{2}{3} + \left\{\frac{n}{4}\right\}\frac{1}{6} = \frac{4\left\lfloor\frac{n}{4}\right\rfloor + \left\{\frac{n}{4}\right\}}{6} = \frac{n}{6}.$$

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

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

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

[Ben-David, 2015] Ben-David, S. (2015). Computational feasibility of clustering under clusterability assumptions. *CoRR*, abs/1501.00437.

[Grønlund et al., 2017] Grønlund, A., Larsen, K. G., Mathiasen, A., and Nielsen, J. S. (2017). Fast exact k-means, k-medians and bregman divergence clustering in 1d. *CoRR*, abs/1701.07204.

[Guruswami and Indyk, 2003] Guruswami, V. and Indyk, P. (2003). Embeddings and non-approximability of geometric problems. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, pages 537–538, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

[Har-Peled and Mazumdar, 2004] Har-Peled, S. and Mazumdar, S. (2004). On coresets for k-means and k-median clustering. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pages 291–300, New York, NY, USA. ACM.

[Khachai and Neznakhina, 2017] Khachai, M. and Neznakhina, E. (2017). Solvability of the Generalized Traveling Salesman Problem in the class of quasi- and pseudo-pyramidal tours (in Russian). *Trudy Inst. Matematiki i Mechaniki UrO RAN*, 23(3):280–291.

[Kumar et al., 2010] Kumar, A., Sabharwal, Y., and Sen, S. (2010). Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM*, 57(2):5:1–5:32.