# Greedy versus Optimum for the $k$-Median Problem

Tomás Feder
268 Waverley St., Palo Alto, CA 94301, USA
tomas@theory.stanford.edu,
Shayan Ehsani, shayan.ehsani@gmail.com,
Amin Saberi
Management Science and Engineering Department, Stanford University,
Stanford, CA 94305, USA
saberi@stanford.edu

April 12, 2012

## 1 Introduction

In the $k$-median problem, we are given $n$ points with positive distances between them satisfying the triangle inequality, and ask to choose $k$ centers out of these $n$ points so that if $d_i$ is the distance to the closest center for point $v_i$, we minimize $T = \sum_{1 \leq i \leq n} d_i$.

This problem is NP-complete but has good polynomial time approximation algorithms, that simultaneously approximate $k$ and $T$ by a factor of $1+\epsilon$ and $2(1+1/\epsilon)$ respectively, for any constant $\epsilon > 0$, see [1]. Also for the exact $k$ a 5-approximation in polynomial time exists, by a local search heuristic that does switches of $k$ centers, which is improved to a $3 + 2/p$ approximation by doing $p$ simultaneous switches, see [2].

In this paper we consider the greedy algorithm that repeatedly selects centers so as to keep reducing the value of $T$ as much as possible, and we ask how many centers need to be chosen to approximate, say within a constant factor, the value of the optimal solution. We first show than on the real line, we need to select at least $2k-1$ centers as a lower bound to get a constant factor approximation to the $k$-median problem. We then show that if $k = 2$, it suffices to select 3 centers to get a constant factor approximation, and $\log_2(1/\epsilon) + O(1)$ centers to get a $1+\epsilon$ approximation to the $k$-median problem. Finally, for any $k$, if $T_r$ is the optimum solution for $r$ centers, we show that $k \log(T_1/(\epsilon T_k))$ centers suffice to get a $1+\epsilon$ approximation to the $k$-median problem.

## 2 $2k-1$ Lower Bound on the Real Line

In this section we consider the case where the points are on the real line.

**Theorem 2.1** *On the real line, we need to select greedily at least $2k-1$ centers to get a constant factor approximation the the $k$-median problem.*

**Proof.** Let $n = 2^{k+\ell} - 1$ for a large constant $\ell$, for each $1 \leq i \leq k$ place $\lfloor n/2^i \rfloor$ points $P_i$ at $\ell^{k-i}$, and place one point $q_i$ for each $1 \leq i < k$ exactly in the middle of $P_i$ and $P_{i+1}$. The gredy algorithm will repeatedly select the single centers at $q_i$ for $i = 1, 2, \ldots, k-1$ as each $q_i$ has exactly the same number of points to the left and to the right before $q_{i+1}$. The other $k$ grouped point sets

$P_i$ will be selected later, for a total of $2k-1$ points selected. The optimum for $k$ selects just the $P_i$. This proves the $2k-1$ lower bound to get a constant factor from the optimum for $k$. $\square$

# 3 The Case $k = 2$

In this case we consider the case where the number of centers being approximated is $k = 2$.

**Theorem 3.1** *If $k = 2$, it suffices to select 3 centers to get a constant factor approximation, and $\log_2(1/\epsilon) + O(1)$ centers to get a $1+\epsilon$ approximation to the $k$-median problem.*

**Proof.** Suppose the two centers of the optimal solution are $C_1, C_2$, the corresponding clusters have $n_1, n_2$, points, where $n = n_1 + n_2$, and the average distance in the clusters is $d_1, d_2$, respectively, so that the optimal solution costs $OPT = n_1 d_1 + n_2 d_2$.

Let $R = \max(n_1 d_1, n_2 d_2)$, let $D_1 = R/n_1, D_2 = R/n_2$, and $OPT' = 2T = n_1 D_1 + n_2 D_2$, with $OPT \leq OPT' \leq 2OPT$. Say $D_1 \leq D_2$ and $n_1 \geq n_2$.

The first center chosen will be at distance $d_1'$ from $C_1$, and at distance $d_2'$ from $C_2$. Let $d_0 = d(C_1, C_2)$. Let $D_1' = \max(d_1', D_1)$ and $D_2' = \max(d_2', D_2)$. Let $D_1' + D_2' = (1+r)d_0$. Then the solution obtained with the first center chosen will have

$$\begin{aligned} S &\geq (D_1' - d_1)n_1 + (D_2' - d_2)n_2 \\ &\geq D_1' n_1 + D_2' n_2 - OPT \\ &\geq ((D_1'/D_1) + (D_2'/D_2))R - OPT \\ &\geq ((D_1'/D_1) + ((d_0(1+r) - D_1')/D_2))R - OPT \\ &\geq d_0(1+r)R/D_2 - OPT. \end{aligned}$$

But by placing the first center at $C_1$ we get

$$OPT + d_0 T/D_2 \geq S \geq d_0(1+r)R/D_2 - OPT$$

$$4R \geq 2OPT \geq d_0 r R/D_2$$

$$r \leq 4D_2/d_0$$

$$D_1' + D_2' \leq d_0 + 4D_2.$$

Choose a large constant $c$. If $d_1' \leq cD_1$ or $d_2' \leq c2$, then we could place the second center at $C_2$, resp. $C_1$, and obtain a $(c+2)$-approximation.

So assume instead $d_1' > cD_1$ and $d_2' > cD_2$. Suppose $D_2 > 2D_1$. Then

$$D_1'(1/D_1 - 1/D_2)R > (D_1'/2D_1)R > (c/2)R \geq (c/4)OPT,$$

giving

$$OPT + d_0 R/D_2 \geq S > d_0(1+r)R/D_2 + (c/4 - 1)OPT,$$

a contradiction.

Finally assume instead $d_1' > cD_1$, $d_2' > cD_2$, and $D_1 \leq D_2 \leq 2D_1$. But now if we place a second center at one of $C_1$ or $C_2$, whichever is better, we go to solution $S'$ with $S' - OPT \leq (S - OPT)/2$, since both $C_1$ and $C_2$ would give OPT or less.

The only way to achieve the half in $S' - OPT \leq (S - OPT)/2$ is essentially a gain in the points of the cluster of $C_1$ or $C_2$, but not both, since $D_1' + D_2' \leq d_0 + 4D_2$ and $4D_2 < 4d_2'/c$. But then the gain is essentially as playing close to $C_1$ or $C_2$, and the third point will give us the $O(1)$ approximation.

For more details, drop first $D_1 \leq D_2$ and $n_1 \geq n_2$, and let $D_3 = \max(D_1, D_2)$. Notice that the second point will be at distance at least $D_1' - 2D_3$ from the center of the cluster of $C_1$, or at least $D_2' - 2D_3$ from the center of the cluster of $C_2$. Assume the first case. Then what used to be a total for cluster of $C_1$ at most $(D'1 + d_1)n_1 \leq (1 + 1/c)D_1'n_1$ becomes at least $(D_1' - 2D_3 - d_1)n_1 \geq (1 - 5/c)D_1'n_1$. But this improvement is a ratio of about $1 - 7/c$, which is minimal since $c$ is large.

It follows that the gain of about half in $S' - OPT \leq (S - OPT)/2$ must be near $C_2$, in particular, the second center must be at small distance $x$ from the center $C_2$. We have

$$(x - d_2)n_2 - d_2n_2 \leq ((d_2' + d_2)n_2 - d_2n_2)/2$$

$$xn_2 \leq d_2'n_2/2 + 2d_2n_2$$

$$x \leq 3d_2'/5.$$

But then for a point is at distance $y$ from $C_1$, the first center chosen had it at distance at most $d_1' + y$, while the center at $x$ has it at distance at least $d_1' + 2d_2'/5 - 4D_3 - y \geq d_1' + d_2'/3 - y$, which can only be an improvement if $y \geq d_2'/6$, which makes an improvement with regards to the points with center in $C_1$ for only $(6D_1/d_2')n_1 < (12/c)n_1$ of the points. But if $x > cD_2$, the improvement in moving $x$ closer to $C_2$ is larger, since it makes a difference for at least $n_2/2$ points at distance at most $2D_2$ from $C_2$, and the improvement for these points is at least $x/2$, while the loss for the $(12/c)n_1$ points near $C_1$ is at most $x$. Since $n_1/2 \leq n_2 \leq 2n_1$, it suffices to take $c = 96$. So the gain must be near the center $C_2$, and then the third point can go near the center $C_1$, giving again a $(c + 2)$-approximation, or a 98-approximation.

Finally since each point cuts the difference to OPT by half, $\log_2(1/\epsilon) + O(1)$ points give a $1 + \epsilon$ approximation. □

## 4 The General Case

We finally show the following.

**Theorem 4.1** *For any $k$, if $T_r$ is the optimum solution for $r$ centers, we show that $k \log(T_1/(\epsilon T_k))$ centers suffice to get a $1 + \epsilon$ approximation to the $k$-median problem.*

**Proof.** The first center gives us $T_1$, and since $k$ additional centers would gives us $T_k$, each choice of centers progresses from $S$ to $S'$ as $S' - T_k \leq (1 - 1/k)(S - T_k)$. The result follows since

$$(1 - 1/k)^{k \log(T_1/(\epsilon T_k))} \leq e^{-\log(T_1/(\epsilon T_k))}$$

$$= \epsilon T_k/T_1.$$

□

# References

[1] J-H Lin and J.S. Vitter, Approximation algorithms for geometric median problems.

[2] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Mungala and V. Pandit, Local search heuristics for $k$-median and facility location problems.