Hardness of Facility Location Problems

Karthik Abinav, Thomas Pensyl, Bartosz Rybicki*

December 11, 2014

1 k-MEDIAN

1.1 Problem Statement

Let F be a set of *facilities* and D be a set of *clients*, and d be a distance metric over $F \cap D$. That is, d_{ij} is the distance between facility i and client j. Given a subset $F' \subseteq F$ of open facilities, we say the connection cost of F' is the total distance from each client to its nearest open facility. k-MEDIAN asks for the cheapest such set of size k.

1.2 Hardness of k-MEDIAN

Theorem 1. *It is hard to approximate k-MEDIAN within $1 + \frac{2}{e^{1/c}}$ for any $c < 1$.*

This theorem is equivalent to the $1 - 1/e - \epsilon$ hardness shown in [2] and follows from a standard reduction from SET COVER. The main idea is to take an approximation algorithm for k-median, and use it to obtain a partial set cover. Then repeat this process, again partially covering the remaining items, until everything is covered. A k-MEDIAN approximation ratio of $1 + \frac{2}{e}$ is just enough to obtain a SET COVER approximation of $\ln n$, which is the best possible under reasonable hardness assumptions. Any better approximation to k-MEDIAN violates this.

1.2.1 The reduction algorithm.

Since all our reductions will be variants of this basic one, it will be useful to give a more general lemma for hardness of partial set covers.

Let $I = (X, S)$ be an instance of SET COVER with sets S over elements X. Let B be an algorithm which, given such an instance, outputs a partial cover $T \subseteq S$. Then define algorithm $C(B)$ which invokes B repeatedly to produce a complete set cover for instance I.

*Collaborator from U of Wroclaw, Poland

Algorithm $C(B)$
Input: $I = (X, S)$,

```
SOL ← ∅
X' ← X
while X' ≠ ∅ do
    T ← B(X', S)                              ▷ Run B to get partial cover T.
    SOL ← SOL ∪ T
    X' ← {x ∈ X' | x not covered by T}        ▷ Get residual instance.
return SOL
```

Notice as long as B always adds at least one set, it covers at least one element, so the algorithm halts after at most $|X|$ loops. Also SOL is indeed a valid set cover for I, since any uncovered element would still be in X'.

Lemma 1. *Fix constants $k > 0$, $\kappa > 0$. Suppose each ith call to B in $C(B)$ covers α_i of the currently uncovered elements, using $\beta_i k \le \kappa k$ sets. Then for every pair of constants $0 < c < c' < 1$, at least one of the following is true:*

a) There exists an i such that $\alpha_i < 1 - e^{-\frac{\beta_i}{c}}$, OR

b) $C(B)$ gives a set cover of size at most $c' \ln |X| \cdot k$.

Proof. Suppose (1) is false. Then there exists a $c < 1$ such that for all i, $\alpha_i \ge 1 - e^{-\frac{\beta_i}{c}}$. Let n_t be the number of uncovered elements remaining after step t. Observing that we start with $|X|$ elements, we have

$$n_t = n_{t-1}(1 - \alpha_t) \le n_{t-1} \cdot e^{-\frac{\beta_t}{c}} = \cdots = |X| \textstyle\prod_{i=1}^{t} e^{-\frac{\beta_i}{c}} = |X| e^{-\frac{1}{c}\sum_{i=1}^{t} \beta_i}.$$

Let $\ell + 1$ be the step after which C terminates. Then after step ℓ, there must still remain at least one uncovered element, so

$$1 \le n_\ell \le |X| e^{-\frac{1}{c}\sum_{i=1}^{\ell} \beta_i} \implies \sum_{i=1}^{\ell} \beta_i \le c \ln |X|.$$

Including the last step, the total sets opened by C is at most

$$\sum_{i=1}^{\ell} \beta_i k + \beta_{\ell+1} k \le (c \ln |X| + \kappa) k = \left(c + \frac{\kappa}{\ln |X|} \right) \ln |X| k \le c' \ln |X| k,$$

for any constant $c' \in (c, 1)$ and $|X| > \exp(\frac{\kappa}{c' - c})$. (We may assume $|X|$ is larger than a constant, since otherwise we could just find the optimal set cover by brute force in constant time.) ☐

By Feige's result [1], unless $NP \subseteq DTIME[n^{O(\log \log n)}]$, it is hard to give a set cover of size $c' \ln |X| \cdot OPT_I$, for any $c' < 1$. Thus, if B never uses more than a constant times OPT_I sets, then Lemma 1 implies it is equally hard for B to give a set which is 'too efficient' in covering. We will define such an algorithm which generates a partial cover by running k-MEDIAN algorithm A on an auxillary facility location instance.

1.2.2 Reduction instance between k-MEDIAN and MAX SET COVER

Let $\Phi_k(I)$ be an instance of k-MEDIAN, using S as the facilities, and X as the clients. We overload S and $s \in S$ to refer to both the sets in I and the facilities in $\Phi_k(I)$, since there is an obvious 1-to-1 equivalence, and the meaning is clear from context; and similarly for X and $x \in X$. For each pair of client $x \in X$ and facility $s \in S$, let the distance between them be 1 if $x \in s$, or 3 if $x \notin s$.

Lemma 2. *Given* SET COVER *instance $I = (X, S)$, let $T \subseteq S$ be a partial set cover. Then T covers exactly $\alpha|X|$ elements in X iff facilities T have client connection cost exactly $(3 - 2\alpha)|X|$.*

Furthermore, if $k = |T| = OPT_I$, then facilities T give a solution to $\Phi_k(I)$ with cost exactly $(3 - 2\alpha)OPT_{\Phi_k(I)}$.

Proof. By definition of $\Phi_k(I)$, each covered element yields a client of cost 1 and each uncovered element yields a client of cost 3. So the total cost is $\alpha|X| \cdot 1 + (1 - \alpha)|X| \cdot 3 = (3 - 2\alpha)|X|$. Furthermore, if $k = |T| = |OPT_I|$, then OPT_I yields a set of k facilities which lets every client pay its lowest possible cost of 1, so $OPT_{\Phi_k(I)} = |X|$. $\square$

Now given a k-MEDIAN algorithm A, define algorithm $B_k(A)$ to take set cover instance I as input and return $A(\Phi_k(I))$. Clearly, $B_k(A)$ returns exactly k sets. Consider $C(B_k(A))$. By trying all possible values of k, we know that in at least one run, $k = OPT_I$, for any input I. By Lemma 1, for any $c < 1$, there must exist an input $I' = (X', S)$ for which $B_k(A)$ covers less than $(1-e^{-1/c})|X'|$ elements. By Lemma 2, this means $A(\Phi_k(I'))$ gives cost strictly greater than

$$(3 - 2(1 - e^{-1/c}))OPT_{\Phi_k(I')} = (1 + \frac{2}{e^{1/c}})OPT_{\Phi_k(I')}.$$

This holds regardless of which algorithm A is used, and thus proves Theorem 1.

1.3 Hardness of r-FAULT TOLERANT k-MEDIAN PLACEMENT

FAULT TOLERANT k-MEDIAN is a generalization of k-MEDIAN in which each client j must pay the connection costs for connecting to r_j facilities, instead of just 1. A special case of this problem is FAULT TOLERANT k-MEDIAN PLACEMENT, in which each facility has $\max_j r_j$ identical, independent copies (i.e. can be opened as many times as is possibly useful). Consider the further specific case where all r_j are uniform, that is, all $r_j = r$ for some r. We denote this problem r-FAULT TOLERANT k-MEDIAN PLACEMENT (rFTkMP). In the case that all $r = 1$, this problem is equivalent to k-MEDIAN, and thus inherits the same $(1 + 2/e)$-hardness of approximation. The nonuniform generalizations also inherit this hardness, which is the best known. However, for the r-uniform versions with $r \geq 2$, including rFTkMP, this hardness does not carry over. We will show how to adapt the standard k-MEDIAN reduction to show a hardness bound as a function of r.

Theorem 2. *It is hard to approximate r-FAULT TOLERANT k-MEDIAN PLACEMENT within $1 + \frac{2}{e^{r/c}}$ for any $c < 1$.*

Define a new reduction instance $\Phi_{\ell,r}(I)$, again using elements X as clients and sets S as facilities. To make it an rFTkP instance, we allow all facilities to be opened any number of times, and set all client demands $r_j = r$. Also set $k = \ell r$. Note when translating facilities back to sets, we allow choosing a set multiple times; this is okay as it does not improve any set cover.

In this case, it is important not just whether an element is covered, but *how many* facilities/sets cover it. To measure this, we define the *coverage vector* $\alpha(T)$ of a multiset T. For $i = 0 \ldots r-1$, let α_i be fraction of elements covered exactly i times. Let α_r be the fraction of elements covered r *or more* times. The following lemma relates the connection cost to this coverage vector.

Lemma 3. *Given* SET COVER *instance $I = (X, S)$, let T be multiset of sets in S, and let $\boldsymbol{\alpha}$ be the coverage vector of T. Also suppose $\ell = OPT_I$. Then facilities T in instance $\Phi_{\ell,r}(I)$ have client connection cost exactly*

$$(3 - 2\sum_{i=1}^{r} \frac{i}{r}\alpha_i) OPT_{\Phi_{\ell,r}(I)}.$$

Proof. Each element requires r connections. If it is covered i times, it has i connections of cost 1 and $r - i$ connections of cost 3. Also note $\sum_{i=0}^{r} \alpha_r = 1$. Then the total cost of all elements is

$$\sum_{i=0}^{r} (3r - 2i)\alpha_i |X| = (3 - 2\sum_{i=1}^{r} \frac{i}{r}\alpha_i) r|X|.$$

Furthermore, if $\ell = |OPT_I|$, then taking r copies of OPT_I yields a set of $k = \ell r$ facilities which lets every client pay its lowest possible cost of r, so $OPT_{\Phi_{\ell,r}(I)} = r|X|$. □

Now given an rFTkMP algorithm A, define algorithm $B_{\ell,r}(A)$ to take set cover instance I as input and return $A(\Phi_{\ell,r}(I))$. Clearly, $B_{\ell,r}(A)$ returns exactly $k = \ell r$ sets. Consider $C(B_{\ell,r}(A))$. By trying all possible values of ℓ, we know that in at least one run, $\ell = OPT_I$, for any particular input I. Now we can apply Lemma 1 with $\kappa = r$: for any $c < 1$, there must exist an input $I' = (X', S)$ for which $B_{\ell,r}(A)$ covers less than $(1 - e^{-r/c})|X'|$ elements. In other words, if $\boldsymbol{\alpha}$ is the coverage vector of the output, we have the number of uncovered elements $\alpha_0 > e^{-r/c}$. By Lemma 3, $A(\Phi_{\ell,r}(I'))$ gives approximation ratio

$$(3 - 2\sum_{i=1}^{r} \frac{i}{r}\alpha_i) \geq (3 - 2\sum_{i=1}^{r} \frac{r}{r}\alpha_i) = (3 - 2(1 - \alpha_0)) = 1 + 2\alpha_0 > 1 + 2^{r/c} \quad (1)$$

where we have used $\sum_{i=0}^{r} \alpha_i = 1$ and $\alpha_0 > e^{-r/c}$. This holds regardless of which algorithm A is used, and thus proves Theorem 2

1.4 Improved bound

Lemma 1 shows it is essentially hard to cover more than $1 - 1/e^r$ elements at least once using $rOPT_I$ sets. However, for (1) to be tight, it must still be easy to cover exactly $1-1/e^r$ elements at least r times each. But can the latter really be easy if the former is hard, especially for large values of r? In this section we will show a negative answer, yielding a tighter hardness result.

Theorem 3. *It is hard to approximate r-FAULT TOLERANT k-MEDIAN PLACEMENT within*

$$\min_{0 \le \gamma \le 1} \left\{ 3 - 2\frac{1 - e^{-\gamma r/c}}{1 - (1-\gamma)^r} \right\}$$

for any $c < 1$.

The idea is this: if a partial cover T really covers its elements many times, than by randomly sampling a subset of T, we can get a partial cover which is much smaller, but still covers a lot of elements at least once. Note that Feige's result holds even for randomized algorithms, so we do not weaken the result by using randomness. Consider for some constant γ, that we sample a uniformly random subset $U \subset T$ with $|U| = \lceil \gamma |T| \rceil$.

Lemma 4. *Given a multiset T of sets over elements X with coverage vector $\boldsymbol{\alpha}$, let $U \subseteq T$ be a random subset of size $\lceil \gamma T \rceil$, chosen uniformly among all such subsets. Then the expected fraction of elements covered by U is at least*

$$\sum_{i=0}^{r} (1-\gamma)^i \alpha_i$$

Proof. The probability that any set $s \in T$ is also in U is at least γ. Let $c_U(x)$ and $c_T(x)$ be the number of sets in U and T which cover element x. The expected number of elements *not* covered by U is

$$\begin{aligned}
E[|\{x \in X \mid c_U(x) = 0\}|] &= \sum_{x \in X} \Pr[c_U(x) = 0] \\
&= \sum_{i=0}^{r} \sum_{x : c_T(x) = i} \Pr[c_U(x) = 0] \\
&= \sum_{i=0}^{r} \sum_{x : c_T(x) = i} \Pr[\text{some } i \text{ sets were not chosen}] \\
&\le \sum_{i=0}^{r} \sum_{x : c_T(x) = i} (1-\gamma)^i \\
&= \sum_{i=0}^{r} (1-\gamma)^i \alpha_i |X|
\end{aligned}$$

□

Define $\Theta(T, \gamma)$ to be a random algorithm which repeats this process some large polynomial number of times, and returns the solution which covers the most elements. Thus, we may assume w.h.p. that the coverage of this solution is at least that of its expected coverage.

Again, given some rFTkMP algorithm A, we define a helper algorithm B.

```
Algorithm B_{ℓ,r}(A):
Input: I = (X, S), c ∈ (0, 1)
--------------------------------------------------------------
  T ← A(Φ_{ℓ,r}(I))
  if |{x ∈ X | x covered by T}| < (1 − 1/e^{r/c})|X| then
      return Θ(T, γ_r)                ▷ Return a random subset of T
  else
      return T
```

Lemma 5. *For any $c < 1$, there is an instance I' for which $B_{\ell,r}(A)$ returns a set with coverage vector $\boldsymbol{\alpha} \in [0,1]^r$ satisfying the following constraints:*

(i) $\sum_{i=0}^{r} \alpha_i = 1$

(ii) $\alpha_0 > e^{-r/c}$

(iii) $\sum_{i=0}^{r} (1 - \gamma_r)^i \alpha_i > e^{-\gamma_r r/c}$

Proof. Consider $C(B_{\ell,r}(A))$. As before, we may try all possible values of ℓ such that in at least one run $\ell = OPT_I$. Lemma 1 (with $\kappa = r$) says that for any $c < 1$, there is an instance I' for which $B_{\ell,r}(A)$ returns $\beta\ell$ sets which cover less than $1 - e^{-\beta/c}$ elements.

Consider $B_{\ell,r}(A)$ when run on instance I'. If T covers at least $1 - e^{-r/c}$ of the elements, then B would return T directly, and the coverage would be too good. This means that T covers less than $1 - e^{-r/c}$ elements. This in turn implies that the returned set is $\Theta(T, \gamma_r)$ (whose coverage is described by Lemma 4) which covers less than $1 - e^{-\gamma_r r/c}$. □

Lemma 3 again says that the corresponding approximation factor obtained by A on $\Phi_{\ell,r}(I')$ is $(3 - 2\sum_{i=0}^{r} \frac{i}{r}\alpha_i)$. This time we minimize it without simplifying as in (1). We find the smallest possible value over all possible $\boldsymbol{\alpha}$ subject to the constraints in Lemma 5. In particular, we have added constraint (iii). For appropriately chosen γ_r, we obtain a stronger lower bound on the approximation factor of A. [1]

[1] It is possible to add arbitrarily many such constraints for many constants γ, but in our calculations we found this did not improve anything

r	1	2	3	4	5	6	7	8
$1+\frac{2}{e^r}$	1.736	1.271	1.010	1.037	1.013	1.005	1.0018	1.0007
new hardness	1.736	1.344	1.224	1.167	1.132	1.110	1.094	1.082
γ_r	1	0.705	0.494	0.379	0.306	0.257	0.222	0.195

Table 1: Calculated values for various r

We claim that the maximum occurs when α_r is maximized such that $\alpha_0 = 1-\alpha_r$, when constraint (iii) is tight.

$$1-\alpha_r+(1-\gamma_r)^r\alpha_r = e^{-\gamma_r r/c}$$

$$\alpha_r = \frac{1-e^{-\gamma_r r/c}}{1-(1-\gamma_r)^r}$$

Then the approximation ratio is

$$3-2\alpha_r = 3-2\frac{1-e^{-\gamma_r r/c}}{1-(1-\gamma_r)^r}$$

Now for each value of r, we will choose parameter γ_r such that this ratio is maximized. As $c \to 1$, we get results shown in Table 1. In particular, we note that the old bound $1+\frac{2}{e^r}$ approaches 1 exponentially fast, while the new bound appears to be close to $1+\frac{2}{re}$, approaching 1 much more slowly.

2 Lower Bounds for Fault Tolerant Facility Placement

To show the lower bound we will use a reduction from set cover. Consider a set cover instance (X, E), where X is the set of sets and E is the set of elements. We will construct an instance of FTFP as follows. For every set $S_i \in X$, have a facility f_i. For every element $e_i \in E$, have a client c_i. We will consider the case of uniform FTFP. All clients have the same demand r. If element $e_i \in S_j$, then add an edge of cost 1 between client c_i and facility f_j. Else, add an edge of cost 3 between c_i and f_j. Let the cost of opening of the factory f_j be q_j. We will define the precise value of q_j a little later. Let $FTFP$ be an α approximation to the FTFP problem. Let k denote the optimal value of the set cover size. We will construct a set cover iteratively as follows

Algorithm for set cover from FTFP
Input: $I = (X, S)$,

$SOL \leftarrow \emptyset$
$X' \leftarrow X$
set cost of facility i as $\gamma_0 * \frac{|E|}{k}$ ▷ γ_0 will be set later
while $X' \neq \emptyset$ **do**
    $T \leftarrow FTFP$ on reduced instance ▷ Run $FTFP$ to get partial cover T.
    $SOL \leftarrow SOL \cup T$
    $X' \leftarrow \{x \in X' \mid \text{x not covered by } T\}$ ▷ Get residual instance.
    set cost of facility i as $\gamma_j * \frac{|X'|}{k}$ ▷ the subscript j refers to the j^{th} round of this loop
return SOL

Consider any j^{th} iteration of the above reduction. Let the cost of the solution returned by FTFP in this round be H. Let us suppose, in this round the number of facilities opened by FTFP is $\beta_j.k$, and let n_j be the number of uncovered elements at the beginning of round j. Also, let c_j be the fraction of clients covered at least once in this round. Hence,

$$H \geq \beta_j.k.\gamma_j.\frac{n_j}{k} + c_j.n_j.r + (1 - c_j).n_j.r.3$$

The first term in the sum refers to cost of opening facilities. The second term refers to the cost of connecting clients who are covered at least once. Note, if they are covered at least once, assuming they are covered r times will give a lower bound. The third term refers to the clients that are not covered even once.

Note that, the optimal solution in this round is

$$\gamma_j.\frac{n_j}{k}.k.r + n_j.r$$

Hence, the value of approximation ratio α should be atleast

$$\frac{\frac{1}{r}.\beta_j.\gamma_j + 3 - 2.c_j}{1 + \gamma_j}$$

We proceed by two cases similar to [3]. Without going into complete details of calculations, we give the results here. The expression is maximised when $\beta_j = c.\ln(\frac{2r}{\gamma_j.c})$. This gives the equation

$$\alpha = \frac{\frac{1}{r}.\gamma_j.\ln(\frac{2r}{\gamma_j}) + 1 + \frac{\gamma_j}{r}}{1 + \gamma_j}$$

Now, we will maximize this with respect to γ_j. Solving this numerically gives us the value of 1.18 for $r = 2$(maximizer $\gamma_j = 0.372$).

References

[1] Feige, Uriel. "A threshold of *lnn* for approximating set cover." *Journal of the ACM (JACM)* 45.4 (1998): 634-652.

[2] Jain, Kamal, Mohammad Mahdian, and Amin Saberi. "A new greedy approach for facility location problems." *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing.* ACM, 2002.

[3] Bartosz Rybicki and Jaroslaw Byrka. Improved approximation algorithm for fault-tolerant facility placement. *CoRR*, abs/1311.6615, 2013