Deterministic Rounding for k-median[1]

- In the k-median problem, we are given a set F of facilities, a set C of clients, and a metric $d(\cdot,\cdot)$ in $F \cup C$. The objective is to open at most k facilities, namely $X \subseteq F$ with $|X| \le k$, and connect clients to the nearest open facilities via assignment $\sigma : C \to X$ so as to minimize

$$\texttt{cost}(X) = \sum_{j \in C} d(\sigma(j), j) \tag{1}$$

- ***LP Relaxation.***

$$\begin{aligned}
\mathsf{lp}(\mathcal{I}) := \text{minimize} \quad & \sum_{i \in F, j \in C} d(i,j) x_{ij} & & \qquad (k\text{-MedLP}) \\
& \sum_{i \in F} x_{ij} = 1, & \forall j \in C & \qquad (2) \\
& y_i - x_{ij} \ge 0, & \forall i \in F,\ \forall j \in C & \qquad (3) \\
& \sum_{i \in F} y_i \le k, & \forall i \in F,\ \forall j \in C & \qquad (4) \\
& x_{ij}, y_i \ge 0, & \forall i \in F, j \in C &
\end{aligned}$$

For every $j \in C$, define $C_j := \sum_{i \in F} d(i,j) x_{ij}$. The rounding algorithm proceeds in phases.

- ***Filtering.*** We consider the clients in *increasing* order of C_j. We add the first client j to a set R. Define $\mathsf{Chld}(j) := \{\ell \in C : d(j,\ell) \le 4\max(C_j, C_\ell)\}$, and remove $\mathsf{Chld}(j)$ from C and continue. At the end of this step, we would have a set of R "representative" clients, and for all $j \in R$, we have a set $\mathsf{Chld}(j) \subseteq C$ clients which partitions C.

procedure FILTERING($F \cup C$, $d(i,j)$):
  Solve (k-MedLP) to obtain (x, y).
  Define $C_j \leftarrow \sum_{i \in F} d(i,j) x_{ij}$.
  $U \leftarrow C$, $R \leftarrow \emptyset$
  while $U \neq \emptyset$ **do**:
    Find $j \in U$ with smallest C_j and $R \leftarrow R \cup j$.
    Set $\mathsf{Chld}(j) \leftarrow \{\ell \in C : d(j,\ell) \le 4\max(C_j, C_\ell)\}$ and $U \leftarrow U \setminus \mathsf{Chld}(j)$.
  return $(R,\ \mathsf{Chld}(j), j \in R)$

Next, for $j \in R$, define $F_j := \{i \in F : d(i,j) \le d(i,k), k \in R\}$. That is, F_j is the subset of the facilities which are closest to j among all representatives, breaking ties arbitrarily. We have the following properties of the filtering procedure.

[1] *Lecture notes by Deeparnab Chakrabarty. Last modified : 3rd Jan, 2022*
These have not gone through scrutiny and may contain errors. If you find any, or have any other comments, please email me at deeparnab@dartmouth.edu. Highly appreciated!

Lemma 1. a. $\mathsf{Chld}(j)\ :\ j \in R$ partitions C and F_j's partition F.
b. For $j \in R$ and $i \in F$, if $d(i,j) \leq 2C_j$, then $i \in F_j$.
c. $y(F_j) := \sum_{i \in F_j} y_i \geq 1/2$ and $|R| \leq 2k$.

Proof. (a) follows from definition of algorithm and F_j's. (b)s is not completely trivial. To see this, suppose not. Suppose there is a facility i with $d(i,j) \leq 2C_j$ but $i \in F_k$ for some $k \neq k$. By definition of F_k, we get $d(i,k) \leq d(i,j) \leq 2C_j$. Triangle inequality implies $d(j,k) \leq 4C_j \leq 4\max(C_j, C_k)$. But then one should have been the child of the other. (c) follows from (b) and an averaging argument: the mass of facilities serving j must be $\geq 1/2$ within a distance twice its contribution to the LP. This implies $|R| \leq 2k$. □

Lemma 2 (Moving clients to the representative)**.** *Given a k-median instance $\mathcal{I}$, consider the k-median instance $\mathcal{I}' = (F, R, \mathsf{dem})$ where on each point $j \in R$, there are $\mathsf{dem}(j) = |\mathsf{Chld}(j)|$ clients co-located. Then, $\mathsf{lp}(\mathcal{I}') \leq \mathsf{lp}(\mathcal{I})$. Furthermore, given any solution $S \subseteq F$ of $|S| \leq k$ facilities, we have $\mathtt{cost}_{\mathcal{I}}(S) \leq \mathtt{cost}_{\mathcal{I}'}(S) + 4\mathsf{lp}(\mathcal{I})$.*

Proof. The first part is by design of R. Given a solution (x,y) to $\mathsf{lp}(\mathcal{I})$, consider the *same* solution for $\mathcal{I}'$ where for every $j' \in \mathsf{Chld}(j)$ we set $x_{ij'} = x_{ij}$ for all $i \in F$. The contribution of $j' \in \mathsf{Chld}(j)$ to $\mathsf{lp}(\mathcal{I}')$ is precisely C_j, and thus is $\leq C_{j'}$, which is the j's contribution to $\mathsf{lp}(\mathcal{I})$.

The second part follows because, by triangle inequality,

$$\mathtt{cost}_{\mathcal{I}}(S) - \mathtt{cost}_{\mathcal{I}'}(S) \leq \sum_{j \in R} \sum_{j' \in \mathsf{Chld}(j)} d(j,j')$$

This is because $d(j',S) \leq d(j',j) + d(j,S)$, and $d(j,S)$ is what j's contribution is to $\mathtt{cost}_{\mathcal{I}'}(S)$. Now, $d(j,j') \leq 2C_j + 2C_{j'} \leq 4C_{j'}$. Summing over all clients, we get that it is $\leq 4\mathsf{lp}(\mathcal{I})$. □

- ***Rounding to a half-integral solution.*** We now consider the instance $\mathcal{I}'$ described in Lemma 2 with co-located clients in $|R|$ distinct positions. Note that the (x,y) solution of $\mathsf{lp}(\mathcal{I})$ is a feasible solution for $\mathcal{I}'$ of cost at most $\mathsf{lp}(\mathcal{I})$. We now massage this fractional solution further potentially increasing the cost, but not by much.

  For every $j \in R$, let $i_1(j) \in F_j$ denote the closest facility to j. Consider a fractional solution where *all* the mass of facilities in F_j is moved to $i_1(j)$. More precisely, define:

$$\forall j \in R, \quad \forall i \in F_j, \quad y'_i := \begin{cases} 0 & \text{if } i \neq i_1(j) \\ y(F_j) \geq \frac{1}{2} & \text{if } i = i_1(j) \end{cases}$$

  Let $F^* := \{i_1(j) : j \in R\}$ be the facilities with any positive y'-mass. Similarly massage the x_{ij}'s as follows

$$\forall j \in R, \forall i \in F, \quad x'_{ij} = \begin{cases} 0 & \text{if } i \notin F^* \\ \sum_{i \in F_k} x_{ij} & \text{if } i = i_1(k) \end{cases} \tag{5}$$

Lemma 3. *(x', y') is a feasible fractional solution for $\mathcal{I}'$ with cost $\mathsf{lp}(x', y') \leq \mathsf{lp}(x,y) + 2\mathsf{lp}(\mathcal{I}')$.*

Proof. Feasibility is easy to see. By design, $\sum_{i \in F} y'_i = \sum_{i \in F} y_i$ since F_j's partition F, and thus (4) is satisfied. For the same reason, for any $j \in R$, we have $\sum_{i \in F} x'_{ij} = \sum_{i \in F} x_{ij}$, and thus (2) is satisfied. $x'_{ij} > 0$ iff $i = i_1(k)$ for some k, and in that case $x'_{ij} \le \sum_{i \in F_k} y_i = y'_i$. Thus, (3) is satisfied.

Let us now consider the *increase* in the cost when one moves from x to x'. Fix a client $j \in R$. The increase in the connection cost of j is

$$\begin{aligned} \sum_{i \in F} d(i,j)\left(x'_{ij} - x_{ij}\right) &= \sum_{k \in R} \sum_{i \in F_k} d(i,j)\left(x'_{ij} - x_{ij}\right) \\ &= \sum_{k \in R} \sum_{i \in F_k} x_{ij} \cdot \left(d(i_1(k), j) - d(i,j)\right) \end{aligned} \tag{6}$$

where we have used (5) for the second equality.

Now, when $k = j$ and for $i \in F_j$, the term $d(i_1(j), j) - d(i,j) \le 0$, by definition of $i_1(j)$. When $k \neq j$, we can still upper bound via the following claim.

Claim 1. *For any $j, k \in R$ and any $i \in F_k$ with $x_{ij} > 0$, $d(i_1(k), j) - d(i,j) \le 2d(i,j)$.*

Note that the claim implies the lemma. To see this, for any client $j \in R$, we can substitute in (6) to get

$$\begin{aligned} \mathsf{lp}(x', y') - \mathsf{lp}(x, y) &= \sum_{j \in R} \mathsf{dem}(j) \cdot \left(\sum_{i \in F} d(i,j)\left(x'_{ij} - x_{ij}\right) \right) \\ &\le \sum_{j \in R} \mathsf{dem}(j) \sum_{i \in F} d(i,j) x_{ij} = 2\mathsf{lp}(\mathcal{I}') \quad \square \end{aligned}$$

- *Proof of Claim 1.* By definition, $d(i_1(k), k) \le d(i,k)$. And, since $i \in F_k$ $d(i,k) \le d(i,j)$. Therefore, by triangle inequality,

$$d(i_1(k), j) \le d(i,j) + d(i,k) + d(i_1(k), k) \le 3d(i,j)$$

- *Moving to $\{\frac{1}{2}, 1\}$-solution.* The fractional solution (x', y') simplifies the picture considerably. There are at most $|R|$ facilities $F^* := \{i_1(j) : j \in R\}$ which have positive y'-value, and furthermore, each $y'_{i_1(j)} \ge \frac{1}{2}$. Note that $i_1(j)$ is the nearest facility to j in F^*. For reasons which will soon be clear, define $i_2(j)$ to be the second nearest facility in F^* to j. Note $i_1(j)$'s are distinct across $j \in R$, but the $i_2(j)$'s may not be distinct.

  Given the y'-values, the best fractional connection cost of any client $j \in R$ is in fact as follows : send $y'_{i_1(j)}$ mass to $i_1(j)$, and send the remaining $(1 - y'_{i_1(j)}) \le \frac{1}{2}$ mass to the $i_2(j)$. Note that this is feasible since $y_{i_2(j)} \ge \frac{1}{2}$. Therefore, we get that

$$\mathsf{lp}(x', y') \ge \sum_{j \in R} \mathsf{dem}(j) \left(d(i_1(j), j) \cdot y_{i_1(j)} \;+\; d(i_2(j), j) \cdot (1 - y_{i_1(j)}) \right) \tag{7}$$

As is, the $y'_{i_1(j)}$'s can be any fraction $\geq 1/2$. However, it is not difficult to massage y''s to $\widehat{y}$'s such that $\widehat{y}_i \in \{\frac{1}{2}, 1\}$ and the RHS of (7) goes down. Indeed, one generic way to see this is to consider the following auxiliary LP with variables $\mathbf{v}$ with v_i for all $i \in F^*$

$$\text{minimize} \quad f(\mathbf{v}) \quad : \quad \sum_{i \in F^*} v_i = k, \;\; 0.5 \leq v_i \leq 1, \forall i \in F^* \tag{8}$$

where $f(\mathbf{v}) = \sum_{j \in R} \mathsf{dem}(j) \left(d(i_1(j), j) \cdot v_{i_1(j)} \; + \; d(i_2(j), j) \cdot (1 - v_{i_1(j)}) \right)$ is a linear function. An extreme point solution must satisfy $|F^*|$ linearly independent inequalities as equality, and since k is an integer this implies $\mathbf{v}_i \in \{0.5, 1\}$. Do you see why? If $\widehat{y}$ is such an extreme point solution, we get $f(\widehat{y}) \leq f(y')$ since y' is a valid solution to the auxiliary LP.

- ***Rounding a $\frac{1}{2}$-integral solution to integral solution.*** Now we are almost done. First, if any $\widehat{y}_i = 1$, we open it. More precisely, let $R' = \{j \in R \; : \; \widehat{y}_{i_1(j)} = \frac{1}{2}\}$; then we open all facilities $\{i_1(j) \; : \; j \in R \setminus R'\}$, and call this set H. We can open $k' := k - |R \setminus R'|$ more facilities. Note that since $\sum_{i \in F^*} \widehat{y}_i = k$, we have that $|R'| = 2k'$.

  For each $j \in R'$, let us draw a directed edge (j, k) from j to $k \in R$ iff $i_2(j) = i_1(k)$. This leads to a directed graph $D = (R, A)$ where every vertex has out-degree at most 1 (vertices in $R \setminus R'$ don't have any out-degree). Thus, D is in fact a collection of *in*-directed trees with possibly one parallel edge with the root. More precisely, each (weakly) connected component is a directed in-tree rooted at some vertex r. All non-root vertices lie in R', and if $r \in R'$, then r has an edge pointing to its child.

  These trees allow us to partition R' into $O \cup E$ by taking the "odd" levels and "even" levels of the tree. This leads to the following property : for all arcs (j, k) if both end points are in R', then one of them is in O and one of them is in E. Now, since $|R'| \leq 2k'$, one of these sets has at most k' clients. Wlog, assume this is O. Then the final k-median algorithm is as follows : for each $j \in O$, open $i_1(j)$ along with the set H facilities opened before.

procedure kMED-ROUNDING($F \cup C$, $d(i,j)$):
    Run FILTERING($F \cup C, d$) to obtain $(R, \mathsf{Chld}(j))$ with $|R| \leq 2k$.
    For all $j \in R$: $F_j \leftarrow \{i \in F : d(i,j) \leq d(i,k), k \in R\}$.
    $F^* \leftarrow \{i_1(j) : j \in R\}$ where $i_1(j)$ is the nearest facility to j in F_j.
    Compute the $\{\frac{1}{2}, 1\}$-solution $\widehat{y}$ given by the solution to (8).
    $H \leftarrow \{i \in F^* \; : \; \widehat{y}_i = 1\}$. ▷ *Let $k' := k - |H|$*
    $R' \leftarrow \{j \in R \; : \; \widehat{y}_{i_1(j)} = \frac{1}{2}\}$ ▷ $|R'| = 2k'$
    For all $j \in R$, $i_2(j)$ is *second* nearest facility to j in F^*.
    Form directed graph $D = (R, A)$ where (j, k) if $i_2(j) = i_1(k)$.
    Using D, partition R' into $O \cup E$ as described above; wlog, $|O| \leq |E|$
    Open $S \leftarrow H \cup \{i_1(j) \; : \; j \in O\}$.

Theorem 1. The algorithm kMED-ROUNDING is a 10-approximation.

Proof. For $j \in R \setminus R'$, it pays $d(i_1(j), j)$ in both the LP and the solution S since $i_1(j) \in H$ is opened. Consider now a $j \in R'$. Note that either $i_1(j)$ is open or $i_2(j)$ is open. Indeed, if $i_1(j) \notin O$, then consider $i_1(k)$ where where $k \in R$ is the unique client with $i_1(k) = i_2(j)$. Either $k \notin R'$ in which case $i_1(k) \in H$ is open. Or, $k \in R'$ which implies $k \in R'$ and thus $i_1(k)$ is open. Therefore, every client $j \in R$ pays at most $\leq d(i_2(j), j)$ in this solution. But in the LP, j pays $\geq \frac{d(i_2(j),j)}{2}$ since $\widehat{y}_{i_1(j)} = 1/2$. Thus, the cost of the algorithm is at most $2 \cdot \mathsf{lp}(x', y')$. By Lemma 3, we get that this cost is $\leq 6\mathsf{lp}(\mathcal{I}') \leq 6\mathsf{lp}(\mathcal{I})$. Since this solution is for $\mathcal{I}'$, porting it to $\mathcal{I}$ and using Lemma 2, we get $\mathtt{cost}(S, \mathcal{I}) \leq 10\mathsf{lp}(\mathcal{I})$. □

Notes

The algorithm described here is the first constant factor approximation algorithm for k-median. This can be found in the paper [2] by Charikar, Guha, Shmoys, and Tardos. That paper consider the special case of $F = C$ and describe a $6\frac{2}{3}$-approximation. Indeed, when $F = C$, the above analysis gives 8-approximation, and we leave the details for the reader. The improvement to 6.67 is obtained by a better rounding of the $1/2$-integral solution to integral (as the reader may have noticed, our analysis has a lot of slack). One can improve the approximation factor of 10 to 8 as is described in the paper [6] by Swamy, but I am not 100% sure if one can go all the way to $6\frac{2}{3}$. My presentation above is borrowed from Swamy's paper. A different randomized rounding algorithm achieving the factor 3.25 can be found in the paper [3], but the analysis is quite involved. The current best approximation factor for k-median is 2.625 which can be found in the paper [1] by Byrka, Pan, Rybicki, Srinivasan, and Trinh. This algorithm however follows a different technique than LP-rounding. It is known that unless $P = NP$, the approximation factor for k-median can't be below 1.735; this result can be found in the papers [4] and [5], respectively. One advantage of the rounding algorithms in [3] and [6] is that they are versatile enough to generalize to capture a host of problems; we refer the reader to Swamy's paper [6] for interesting applications.

References

[1] J. Byrka, T. Pensyl, B. Rybicki, A. Srinivasan, and K. Trinh. An improved approximation for k-median, and positive correlation in budgeted optimization. In *Proc., ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 737–756, 2014.

[2] M. Charikar, S. Guha, D. B. Shmoys, and É. Tardos. A Constant Factor Approximation Algorithm for the k-median Problem. *Proceedings of 31st ACM STOC*, 1999.

[3] M. Charikar and S. Li. A dependent lp-rounding approach for the k-median problem. In *Proc., International Colloquium on Automata, Languages and Programming (ICALP)*, pages 194–205, 2012.

[4] S. Guha and S. Khuller. Greedy Strikes Back: Improved Facility Location Algorithms. In *Proc., ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 649–657, 1998.

[5] K. Jain, M. Mahdian, and M. Salavatipour. Packing Steiner Trees. *Proceedings of* 13^{th} *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 266–274, 2003.

[6] C. Swamy. Improved approximation algorithms for matroid and knapsack median problems and applications. *ACM Trans. on Algorithms (TALG)*, 12(4):1–22, 2016.