CMPUT 675: Approximation Algorithms **Winter 2018**

Lecture 11 (Feb 13): k-median via Lagrangian Relaxation

Lecturer: Mohammad R. Salavatipour *Scribe: Dylan Hyatt-Denesik*

11.1 k-median via Lagrandian Relaxation

Recall, for the k-median problem we are given a graph $G = (V, E)$ along with a cost function $c : E \to \mathbb{Q}^+$ satisfying the metric property. We also have a set $F \subseteq V$ of centres and a set $C \subseteq V$ of clients, plus an integer k. The goal is to open/select k centres and assign each client to the nearest selected centre to minimize the sum of distances of clients to their centres. The following IP formulates k-median:

$$\begin{array}{lrcll} \min & \sum_{i\in F, j\in C} c_{ij}x_{ij} & & & \\ \text{s.t.} & \sum_{i\in F} x_{ij} & \geq & 1 & j \in C \\ & y_i - x_{ij} & \geq & 0 & i \in F, j \in C \\ & \sum_{i\in F} -y_i & \geq & -k & \\ & x_{ij}, y_i \in \{0,1\}, & & & i \in F, j \in C \end{array} \tag{11.1}$$

By relaxing the last constraint to $0 \leq x_{ij} \leq 1$ obtain an LP. Define dual variables α for the first constraint, β for the second constraint , and z for the last. We then find the following dual LP

$$\begin{array}{lrcll} \max & \sum_{j\in C} \alpha_j - zk & & & \\ \text{s.t.} & \alpha_j - \beta_{ij} & \leq & c_{ij} & i \in F, j \in C \\ & \sum_{i\in F} \beta_{ij} - z & \leq & 0 & i \in F \\ & \alpha_j, \beta_{ij}, z \geq 0, & & & i \in F, j \in C \end{array} \tag{11.2}$$

11.1.1 The high level idea

The linear programs above and those for uncapacitated facility location are very similar and we will make use of this similarity to come up with an algorithm to solve the k-median problem. To make use of this similarity we assign a cost of $f_i = z$ for opening each facility $i \in F$. Suppose we know a good value for z such that if we run the facility location primal/dual algorithm it opens exactly k centres. In other words we find primal and dual solutions $(\mathbf{x}, \mathbf{y})$ and (α, β) respectively, where

$$\sum_i y_i = k \qquad \text{and} \qquad \sum_{\mathrm{i\in F, j\in C}} \mathrm{c_{ij}x_{ij}} + \mathrm{3zk} \leq 3\sum_{\mathrm{j\in C}} \alpha_{\mathrm{j}}$$

Therefore:

$$\sum_{i\in F, j\in C} c_{ij}x_{ij} \leq 3(\sum_{j\in C} \alpha_j - zk)$$

Hence, we find an integral solution for k-median that is in 3 times the optimal. However, there is a problem with this approach: what is the value for z? In the case the algorithm opens more than k facilities, then the solution is infeasible for the k-median problem. If it opens fewer than k facilities, then the above inequalities fail.

Thus, the problem is to find z such that the algorithm opens exactly k facilities. Observe, if $z = 0$ then the algorithm will open all $|F|$ facilities, and if z is a sufficiently large value M then the algorithm opens only 1 facility. So, if c_{max} is the largest client cost, we can perform binary search on the range $[0, nc_{max}]$ to find values z_1 and z_2 such that the algorithm opens $k_1 > k$ facilities and $k_2 < k$ facilities respectively, and such that $z_1 - z_2 \leq \frac{c_{min}}{O(n^2)}$, where $n_c = |C|$.

We let (x^1, y^1) and (x^2, y^2) be the primal solutions induced from the above values of z, and (α^1, β^1) and (α^2, β^2) be the corresponding dual solutions. Pick a and b such that $k = ak_1 + bk_2$, and define $(x, y) = a(x^1, y^1) + b(x^2, y^2)$, note $a = (k_2 - k)/(k_2 - k_1)$ and $b = (k - k_1)/(k_2 - k_1)$. It is not hard to see that these convex solutions satistfy the constraints of (11.1) and (11.2). To see this, notice that the left side of each constraint can be expanded into a convex combination of the solutions for z_1 and z_2, which will be bounded by a convex combination of the right hand side of the constraint which will just be the constraint.

Lemma 1 *The cost of (x, y) is within a factor of $(3 + 1/n_c)$ of the cost of an optimal fractional solution OPT of the k-median problem*

Proof. Recall the following inequalities for feasible solutions

$$\sum_{i \in F, j \in C} c_{ij} x^1_{ij} \leq 3(\sum_{j \in C} \alpha^1_j - z_1 k_1) \tag{11.3}$$

$$\sum_{i \in F, j \in C} c_{ij} x^2_{ij} \leq 3(\sum_{j \in C} \alpha^2_j - z_2 k_2) \quad (2) \tag{11.4}$$

We will also need to make note of the following two facts

- Fact 1: $z_1 - z_2 \leq c_{min}/O(n^2)$
- Fact 2: $\sum_{ij} c_{ij} x^1_{ij} \geq n_c c_{min}$

Since $z_1 > z_2$ we have that (α^2, β^2) is a feasible dual solution to the facility location problem; even if the facilities have a cost of z_1. To prove the desired bound we will find an updated bound for inequality (11.2). Using above facts we can find:

$$\sum_{i \in F, j \in C} c_{ij} x^2_{ij} \leq 3(\sum_{j \in C} \alpha^2_j - z_2 k_2) = 3(\sum_{j \in C} \alpha^2_j - (z_2 + z_1 - z_1)k_2) = 3(\sum_{j \in C} \alpha^2_j - z_1 k_2) + 3(z_1 - z_2)k_2$$

$$\leq 3(\sum_{j \in C} \alpha^2_j - z_1 k_2) + k_2 c_{min}/O(n^2) \leq 3(\sum_{j \in C} \alpha^2_j - z_1 k_2) + OPT/n_c$$

Where the first and second inequalities follow from Facts 1 and 2 respectively, the third inequality follows since $k_2 < n$ and $c_{min} \leq OPT$.

Adding the inequality that we find above, multiplied by b with inequality (11.1) multiplied by a, we easily find the following:

$$\sum_{i \in F, j \in C} c_{ij} x_{ij} \leq a(3(\sum_{j \in C} \alpha^1_j - z_1 k_1)) + b(3(\sum_{j \in C} \alpha^2_j - z_1 k_2) + OPT/n_c)$$

$$= 3(\sum_{j \in C} \alpha_j^1 - z_1 k) + bOPT/n_c \leq (3 + 1/n_c)OPT$$

Where $\alpha = a\alpha^1 + b\alpha^2$ and $\beta = a\beta^1 + b\beta^2$. All that remains to be seen is the fact that (α, β, z_1) is a feasible solution to the k-median dual. This certainly holds since (α^1, β^1) and (α^2, β^2) are feasible, thus $\sum_{j \in C} \beta_{ij}^l \leq z_l$ for $l \in \{1, 2\}$. So $\sum_{j \in C} \beta_{ij} \leq z = az_1 + bz_2 \leq (a + b)z_1 = z_1$. So the last inequality holds. ■

11.2 Obtaining an integral solution to the k-Median

In this section we give a randomized rounding procedure for the k-median problem which rounds the solution from the previous section, and in doing so will only increase the cost by a factor of $1 + \max(a, b)$. Let A and B be the sets of facilities opened by solutions (x^1, y^1) and (x^2, y^2) respectively, so $|A| = k_1$ and $|B| = k_2$. The rounding procedure is given as follows:

For each facility in A, find the nearest facility in B, and denote these facilities by B'. If $|B'| < k_1$, then arbitrarily include facilities from $B - B'$ until $|B'| = k_1$. With probability a, open all facilities in A, and with probability $b = 1 - a$, open the facilities in B'. Then with uniform, random probability pick a set of $k - k_1$ facilities from $B - B'$, and open these as well. Denote the set of facilities opened by I.

Let $\phi : C \to I$ be the assignment of clients to facilities in the randomized solution. For $j \in C$ suppose $i_1 \in A$ and $i_2 \in B$ are the facilities j is assigned to in both solutions, then we have the following cases:

- Case 1: $i_2 \in B'$. then one of i_1 and i_2 is opened by the procedure above, with probability a and b respectively. j is simply assigned to the open one.
- Case 2: $i_2 \notin B'$. Let $i_3 \in B'$ be the facility nearest to i_1. If i_2 is open, then j is assigned to it. If $i_2 \in I$ then j assigned to i_2, note that this happens with probability b. Otherwise, assign j to i_1 if i_1 is open. In the case that neither is open, assign j to i_3.

Denote $c_j^* = ac_{i_1 j} + bc_{i_2 j}$, the cost of assigning client j in the convex solution (x, y).

Lemma 2 *For each client $j \in C$, $\mathbb{E}[c_{\phi(j)j}] \leq (1 + \max(a, b))c^*$*

Proof. To prove this lemma we examine the cases for ϕ. If $i_2 \in B'$, then $\mathbb{E}[c_{\phi(j)j}] = ac_{i_1 j} + bc_{i_2 j} = c_j^*$

If $i_2 \notin B'$, then i_2 is open with probability b. The probability i_2 is not open and whereas i_1 is open is $(1-b)a = a^2$, the probability both are not open is $(1 - b)(1 - a) = ab$. Thus we have $\mathbb{E}[c_{\phi(j)j}] \leq bc_{i_2 j} + a^2 c_{i_1 j} + abc_{i_3 j}$.

Since i_3 is in the facility in B that is closest to i_1 we have by the triangle inequality, $c_{i_1 i_3} \leq c_{i_1 i_2} \leq c_{i_1 j} + c_{i_2 j}$, similarly we have $c_{i_3 j} \leq c_{i_1 j} + a^2 c_{i_1 j} \leq 2c_{i_1 j} + c_{i_2 j}$. Noting $a^2 c_{i_1 j} + abc_{i_1 j} = ac_{i_1 j}$ we find

$$\mathbb{E}[c_{\phi(j)j}] \leq bc_{i_2 j} + a^2 c_{i_1 j} + ab(2c_{i_1 j} + c_{i_2 j}) \leq (ac_{i_1 j} + bc_{i_2 j}) + ab(c_{i_1 j} + c_{i_2 j})$$

$$\leq (ac_{i_1 j} + bc_{i_2 j})(1 + \max(a, b)) = (1 + \max(a, b))c_j^*$$

■

Let (x^k, y^k) denote the integral solution found by this randomzied procedure. We can easily see using Lemma 2 that,

$$\mathbb{E}[\sum_{i \in F, j \in C} c_{ij} x_{ij}^k] \leq (1 + \max(a, b)) \sum_{i \in F, j \in C} c_{ij} x_{ij} \leq 2 \sum_{i \in F, j \in C} c_{ij} x_{ij}$$

Combinnig Lemma 1 and Lemma 2 we finally observe the following theorem.

Theorem 1 *Using the primal/dual algorithm for Facility Location and Lagrangian Relaxation for the k-Median problem, we find a* $(6+\varepsilon)$*-approximation to the k-Median problem.*