Assignment Week 3: A Primal-Dual Algorithm for the k-median Problem.

In this exercise, we propose to design a primal-dual algorithm for the k-median problem.

The k-median problem is defined as follows: given a complete bipartite graph $G = (F \cup C, E)$ where F is the center set and C the client set and a metric distance function $c : E \to \mathbb{R}^+$ the goal is to find a subset $S \subseteq F$ containing at most k elements and such that $\sum_{v \in C} \min_{f \in S} c((v, f))$ is minimized.

This problem is similar to the facility location problem, except that there is no cost for opening a facility.

We want to design a primal-dual algorithm achieving a 6-approximation based on the primal dual algorithm seen during the lectures.

The standard LP formulation, LP1, is as follows:

minimize

$$\sum_{j \in C} \min_{i \in F} x_{ij} c((i, j)). \tag{1}$$

subject to

$$\begin{aligned} \forall j \in C, \quad & \sum_{i \in F} x_{ij} \geq 1\,, \\ \forall j \in C,\, i \in F, \quad & y_i - x_{ij} \geq 0\,, \\ & \sum_{i \in F} y_i \leq k\,, \\ \forall j \in C,\, i \in F\,, \quad & x_{ij} \geq 0\,, \\ \forall i \in F\,, \quad & y_i \geq 0\,. \end{aligned} \tag{2}$$

In order to have a LP, LP2, that is closer to the one we had for facility location, we will consider a slightly different LP. For a given $\lambda \geq 0$, we define the following LP.

minimize $\sum_{j \in C} \min_{i \in F} x_{ij} c((i, j)) + \lambda(\sum_{i \in F} y_i - k)$.

subject to

$$\begin{aligned} \forall j \in C, \quad & \sum_{i \in F} x_{ij} \geq 1\,, \\ \forall j \in C,\, i \in F, \quad & y_i - x_{ij} \geq 0\,, \\ \forall j \in C,\, i \in F\,, \quad & x_{ij} \geq 0\,, \\ \forall i \in F\,, \quad & y_i \geq 0\,. \end{aligned} \tag{3}$$

The parameter λ is a way to penalize the LP for opening a lot of facilities. Indeed, we would like to find a good value for λ so that an optimal solution for the LP will open at most k facilities.

Observe that any feasible solution for the linear programming relaxation of the k-median problem, LP1, is also feasible for LP2.

Moreover, any feasible solution for LP2 is also a solution for LP1, assuming $\lambda \geq 0$.

- Question 1: Give the dual of LP2, DUAL2.

  The dual has variables α_j, β_{ij} $(i \in F, j \in C)$. The goal is:

$$\max_{j \in C} \alpha_j - \lambda k\,, \tag{4}$$

subject to

$$\begin{aligned} \forall j \in C,\, i \in F, \quad & \alpha_j - \beta_{ij} \leq c((i,j))\,, \\ \forall i \in F, \quad & \sum_{j \in C} \beta_{ij} \leq \lambda\,, \\ \forall j \in C\,, \quad & \alpha_j \geq 0\,, \\ \forall j \in C,\, i \in F\,, \quad & \beta_{ij} \geq 0\,. \end{aligned} \tag{5}$$

Note the presence of the constant term $-\lambda k$ in the objective. Since the primal also has such a constant term in the objective we have to add it to the dual to make the weak/strong duality theorem work: the duality theorem works without the extra constant, then we just add the same constant term to both the primal and dual objective.

We now would like to use the primal-dual algorithm seen during the lectures (the one that achieves a 3-approximation) in which all the facility cost are equal to some $\lambda \geq 0$.

Recall that in Lecture 8, we have seen that $\sum_{\text{cluster } C_0} \left(f_{i_{C_0}} + \sum_{j \in C_0} c((i,j)) \right) \leq 3 \sum_{j \in C_0} \alpha_j$.

In fact, it is possible to show that $\sum_{\text{cluster } C_0} \left(3\, f_{i_{C_0}} + \sum_{j \in C_0} c((i,j)) \right) \leq 3 \sum_{j \in C_0} \alpha_j$.

- Question 2: Suppose that the f_i are all equal to λ and that the algorithm opens a set S of facilities, what can you deduce from the above formula?

  We have $\sum_{\text{cluster } C_0} 3\, f_{i_{C_0}} = \sum_{\text{cluster } C_0} 3\,\lambda = 3\,\lambda |S|$ and thus

$$3\,\lambda |S| + \sum_{C_0} \sum_{j \in C_0} c((i,j)) \leq 3 \sum_{j \in S} \alpha_j\,. \tag{6}$$

- Question 3: Suppose further that the set S contains exactly k facilities, what can you deduce from the cost of the solution output by the algorithm and the value of the dual?

  From Q2 and the objective of LP2, $\text{cost}(S) = \sum_{C_0} \sum_{j \in C_0} c((i,j)) + \lambda(|S| - k)$, we find

$$2\,\lambda(|S| - k) + \text{cost}(S) \leq 3 \left(\sum_{j \in S} \alpha_j - \lambda k \right) = 3\,\text{val}(\alpha) \leq 3\,\text{OPT}\,, \tag{7}$$

  and putting $|S| = k$

$$\text{cost}(S) \leq 3 \left(\sum_{j \in S} \alpha_j - \lambda k \right) = 3\,\text{val}(\alpha) \leq 3\,\text{OPT} \tag{8}$$

  We now want to find the value of λ that would lead the algorithm to open exactly k facilities.

  In order to do so, we will do a bisection search and maintain a lower bound λ_1 and an upper bound λ_2 on the value of the optimal λ.

  We start with $\lambda_1 = 0$ and $\lambda_2 = \sum_{j \in C} \sum_{i \in F} c((i,j))$.

- Question 4: Prove that in the case of $\lambda = \lambda_1$, the algorithm opens at least k facilities and in the case of $\lambda = \lambda_2$ the algorithm opens less than k facilities.

  If $\lambda = \lambda_1 = 0$, the algorithm blocks for each client j the edge with minimum c_{ij}, i.e. the edge to the closest facility, with $\alpha_j = \min_i c_{ij}$. It also blocks this closest facility. Assuming all edge weights are different no client is connected with a minimal edge to two facilities (if there are equal minimal edges we can arbitrarily chose one to be blocked). It follows that there are no clients at distance 3 to a blocked facility and all blocked facilities are opened. So the opened facilities are all the facilities which are closest to at least one edge. If there is no constraint on the number of facilities, this is obviously the optimal solution. We assume the number of such facilities is bigger than k, otherwise the problem is trivial.

If $\lambda = \lambda_2$, the first facility i to be blocked is the one for which $\sum_j c_{ij}$ is minimal. At that point we have $\alpha_j \geq (\lambda_2 + \min_i c_{ij})/|C| \geq c_{ij}$ for all j, all edges are blocked, and all clients are at distance at most 3 of the facility. So only one facility is opened.

We start by running the algorithm for λ_1, which returns a set of facilities S_1. Then we run the algorithm for λ_2 and obtain a set of facilities S_2. If $|S_1| > k$ and $|S_2| < k$, we run the algorithm on the value $\lambda = (\lambda_1 + \lambda_2)/2$.

The algorithm outputs a set S of facilities. If $|S| > k$, we set $\lambda_1 = (\lambda_1 + \lambda_2)/2$, $S_1 = S$ and repeat.

Otherwise we set $\lambda_2 = (\lambda_1 + \lambda_2)/2$, $S_2 = S$ and repeat.

We repeat until we obtain a set S of k facilities or $\lambda_2 - \lambda_1$ is small enough.

In the later case, we explain how to combine S_1 and S_2 to obtain a solution with k facilities.

Let $c_{\min}$ be the smallest assignment cost greater than 0. We run the bisection search until we get a set S of k facilities or until $\lambda_2 - \lambda_1 \leq \epsilon\, c_{\min}/(3|F|)$, for some fixed $\epsilon > 0$.

If we have not terminated with a solution with exactly k facilities, the algorithm terminates with solutions S_1 and S_2 and (corresponding) dual solutions (α^1, β^1) and (α^2, β^2) such that $|S_1| > k > |S_2|$.

- Question 5: Use Question 3 to derive an inequality connecting the value of the solution induced by the set S_1, $\text{cost}(S_1)$, and (α_1, β_1) and λ_1. Derive a similar bound for the value of the solution induced by S_2, $\text{cost}(S_2)$.

  For simplicity let us define the cost without the Lagrange multiplier part

$$\text{cost}'(S) = \sum_{C_0} \sum_{j \in C_0} c((i,j)) = \text{cost}(S) - \lambda(|S| - k)\,. \tag{9}$$

  From Q3 we find then for S_1

$$\text{cost}'(S_1) \leq 3 \sum_j \alpha_j^1 - 3\,\lambda |S_1|\,, \tag{10}$$

  and S_2

$$\text{cost}'(S_2) \leq 3 \sum_j \alpha_j^2 - 3\,\lambda |S_2|\,. \tag{11}$$

  Without loss of generality, we can assume that $0 < c_{\min} \leq \text{OPT}$, since if $\text{OPT} = 0$ then it is easy to compute an optimal solution.

  We now pick $\delta_1, \delta_2 > 0$ such that $\delta_1 + \delta_2 = 1$ and $\delta_1 |S_1| + \delta_2 |S_2| = k$.

  We can then get a dual solution $(\tilde{\alpha}, \tilde{\beta})$ by letting $\tilde{\alpha} = \delta_1 \alpha_1 + \delta_2 \alpha_2$ and $\tilde{\beta} = \delta_1 \beta_1 + \delta_2 \beta_2$.

  Note that $(\tilde{\alpha}, \tilde{\beta})$ is feasible for the DUAL2 with facility costs λ_2 since it is a convex combination of two feasible dual solutions. We can now prove the following lemma.

  It states that the convex combination of the costs of the solutions induced by S_1 and S_2 must be close to the cost of an optimal solution.

  Lemma 1. $\delta_1 \text{cost}(S_1) + \delta_2 \text{cost}(S_2) \leq (3 + \delta_1 \epsilon)\text{OPT}$.

- Question 6: Using Question 5 and the fact that $\lambda_2 - \lambda_1 \leq \epsilon c_{\min}/(3|F|)$, prove that $\text{cost}(S_1) \leq 3\left(\sum_{j \in C} \alpha_j^1 - \lambda_2 |S_1|\right) + \epsilon\,\text{OPT}$.

We find using respectively Q5, $|S_1| > k$, $-\lambda_1 \leq \epsilon c_{\min}/(3|F|) - \lambda_2$, $c_{\min} \leq \mathrm{OPT}$ and $|S_1| \leq |F|$

$$\begin{aligned}
\mathrm{cost}'(S_1) &\leq 3 \sum_j \alpha_j^1 - 3\,\lambda_1 |S_1| \\
&\leq 3 \sum_j \alpha_j^1 - 3\,\lambda_2 |S_1| + \epsilon\, c_{\min} \frac{|S_1|}{|F|} \\
&\leq 3 \left(\sum_j \alpha_j^1 - \lambda_2 \right) + \epsilon\, \mathrm{OPT}\,.
\end{aligned} \tag{12}$$

- Question 7: Using a convex combination of $\mathrm{cost}(S_2)$ and the inequality derived in Question 6, conclude the proof of Lemma 1.

  Combining Q5 and Q6 we find

$$\begin{aligned}
\delta_1\, \mathrm{cost}'(S_1) + \delta_2\, \mathrm{cost}'(S_2) &\leq 3\,\delta_1 \sum_j \alpha_j^1 + 3\,\delta_2 \sum_j \alpha_j^2 - 3\,(\delta_1 + \delta_2)\,\lambda_2 + \delta_1 \epsilon\, \mathrm{OPT} \\
&\leq 3\,\mathrm{val}(\tilde{\alpha}) + \delta_1 \epsilon\, \mathrm{OPT} \\
&\leq (3 + \delta_1 \epsilon)\, \mathrm{OPT}\,.
\end{aligned} \tag{13}$$

  Let us now consider the original costs:

$$\begin{aligned}
\delta_1\, \mathrm{cost}(S_1) + \delta_2\, \mathrm{cost}(S_2) &= \delta_1\, \mathrm{cost}'(S_1) + \delta_2\, \mathrm{cost}'(S_2) + \delta_1 \lambda_1 (|S_1| - k) + \delta_2 \lambda_2 (|S_2| - k) \\
&\leq \delta_1\, \mathrm{cost}'(S_1) + \delta_2\, \mathrm{cost}'(S_2) + \delta_1 \lambda_1 (|S_1| - k) + \delta_2 \lambda_1 (|S_2| - k) \\
&\quad + \epsilon \delta_2 \frac{c_{\min}}{3|F|}\, (|S_2| - k) \\
&\leq 0\,,
\end{aligned} \tag{14}$$

  where we used $\lambda_2 - \lambda_1 \leq \epsilon c_{\min}/(3|F|)$, $\delta_1 |S_1| + \delta_2 |S_2| = k$ and $|S_2| \leq k$. Lemma 1 follows.

  We now conclude the analysis. We need to distinguish two cases, $\delta_2 \geq 1/2$ and $\delta_2 < 1/2$.

  In the case of $\delta_2 \geq 1/2$, we return the set S_2. Recall that $|S_2| < k$ and thus, S_2 is feasible.

- Question 8: Assuming $\delta_2 \geq 1/2$ and using Lemma 1, show that the set S_2 is a solution of cost at most $2(3 + \epsilon)\mathrm{OPT}$.

  From the lemma we find

$$\mathrm{cost}(S_2) \leq 1/\delta_2\, (\delta_1 \mathrm{cost}(S_1) + \delta_2 \mathrm{cost}(S_2)) \leq 1/\delta_2 (3 + \delta_1 \epsilon)\mathrm{OPT} \leq 2(3 + \delta_1 \epsilon)\mathrm{OPT}\,. \tag{15}$$

  We now assume that $\delta_2 < 1/2$.

  Then, for each facility $i \in S_2$, we open the closest facility $h \in S_1$; that is, the facility $h \in S_1$. If this doesn't open $|S_2|$ facilities of S_1 because some facilities in S_2 are close to the same facility in S_1, we open some arbitrary facilities in S_1 so that exactly $|S_2|$ are opened. We then choose a random subset of $k - |S_2|$ of the $|S_1| - |S_2|$ facilities of S_1 remaining, and open these. Let S be the resulting set of facilities opened.

  We show that the expected cost of S is at most $2(3 + \epsilon)\mathrm{OPT}$.

  We give a bound on the expected cost of assigning a given client j to a facility opened by the randomized algorithm.

  Let us suppose that the facility $f_1 \in S_1$ is the open facility in S_1 closest to j. This means that $c((f_1, j))$ is the contribution of client j to the cost of S_1, c_j^1.

  Let $f_2 \in S_2$ be the open facility in S_2 that is the closest to j. Again, $c((f_2, j))$ is the contribution of client j to the cost of S_2, c_j^2.

  Recall that $\frac{k - |S_2|}{|S_1| - |S_2|} = \delta_1$.

- Question 9: What is the probability that the randomized algorithm opens f_1?

  The algorithm opens $k - |S_2|$ facilities of the remaining $|S_1| - |S_2|$, so the probability is $\delta_1 = \frac{k-|S_2|}{|S_1|-|S_2|}$.

  If f_1 is open, then j is assigned to f_1. Otherwise, we assign j to the closest facility of S_1 closest to f_2. Let i be this facility. Recall that by the triangle inequality we have that $c((i,j)) \leq c((j,f_2)) + c((f_2,i))$.

- Question 10: Give an upper bound on the distance of $c((i,f_2))$ based on the distance from $c((f_1,f_2))$

  We have:

  $$c((i,f_2)) \leq c((f_1,f_2)) \tag{16}$$

  because i is the closest facility to f_2 of S_1 and thus closer than f_1.

- Question 11: Show that $c((i,j)) \leq c_j^1 + 2c_j^2$.

  Using the triangle identity and Q10 we find:

  $$c((i,j)) \leq c((j,f_2)) + c((i,f_2)) \leq c((j,f_2)) + c((f_1,f_2)) \leq c((j,f_1)) + 2\,c((j,f_2)) = c_j^1 + 2\,c_j^2\,. \tag{17}$$

- Question 12: Based on Questions 9 and 11, show that the expected cost for client j is at most $\delta_1 c_j^1 + \delta_2(c_j^1 + 2c_j^2)$.

  The probability that f_1 is opened is δ_1 and the cost for client j is then c_j^1, in the other case, with probability $1 - \delta_1 = \delta_2$, the cost is $c((i,j))$.

  We find for the expected cost using Q11:

  $$\delta_1 c_j^1 + \delta_2 c_j^2 \leq \delta_1 c_j^1 + \delta_2(c_j^1 + 2c_j^2)\,. \tag{18}$$

- Question 13: Conclude the proof of this case using Question 12 and the fact that $\delta_2 < 1/2$.

  The cost for each client is bounded by:

  $$\delta_1 c_j^1 + \delta_2(c_j^1 + 2c_j^2) \leq 2\,\left(\delta_1 c_j^1 + \delta_2 c_j^2\right)\,, \tag{19}$$

  and thus using Lemma 1 and the fact that $|S| = k$ by construction (so the Langrange multiplier term vanishes):

  $$\text{cost} \leq 2\sum_j \left(\delta_1 c_j^1 + \delta_2 c_j^2\right) \leq 2\,(3+\epsilon)\text{OPT}\,. \tag{20}$$