Assignment Week 4: An SDP based randomized algorithm for the Correlation Clustering problem

The objective of this exercise is to design an algorithm for the *correlation clustering problem.* Given an undirected graph $G = (V, E)$ without loops, for each edge $e = \{i, j\} \in E$ there are two non-negative numbers $w_e^+, w_e^- \geq 0$ representing how similar and dissimilar the nodes i and j are, respectively. For $S \subseteq V$, let $E(S)$ be the set of edges with both endpoints in S, that is, $E(S) = \{\{i, j\} \in E : i, j \in S\}$. The goal is to find a partition $\mathcal{S}$ of V that maximizes

$$f(\mathcal{S}) = \sum_{S \in \mathcal{S}: e \in E(S)} w_e^+ + \sum_{e \in E \setminus \cup_{S \in \mathcal{S}} E(S)} w_e^- . \tag{1}$$

In words, the objective is to find a partition that maximizes the total similarity inside each set of the partition plus the dissimilarity between nodes in different sets of the partition.

Consider the following simple algorithm:

Algorithm 1

Let $\mathcal{S}_1 = \{\{i\} : i \in V\}$ the partition that considers each vertex as a single cluster, and $\mathcal{S}_2 = \{V\}$, that is every vertex in the same cluster. Compute the values $f(\mathcal{S}_1)$ and $f(\mathcal{S}_2)$ of these two partitions, and output the best among this two.

- Question 1. Compute the values $f(\mathcal{S}_1)$, $f(\mathcal{S}_2)$ in terms of the weights w^- and w^+.

  For $\mathcal{S}_1$ there are no edges in any $E(S)$ with $S = \{i\} \in \mathcal{S}_1$, while for $\mathcal{S}_2$ all edges are in $E(V)$. So we find

$$\begin{aligned} f(\mathcal{S}_1) &= \sum_{e \in E} w_e^- \, , \\ f(\mathcal{S}_2) &= \sum_{e \in E} w_e^+ \, . \end{aligned} \tag{2}$$

- Question 2. Conclude that previous algorithm is a 1/2-approximation.

  We have the following bound for any $\mathcal{S}$:

$$f(\mathcal{S}) \leq \sum_e w_e^+ + \sum_e w_e^- = f(\mathcal{S}_1) + f(\mathcal{S}_2) \leq 2 \max(f(\mathcal{S}_1), f(\mathcal{S}_2)) \, , \tag{3}$$

  and thus also for the optimal $\mathcal{S}^*$

$$\max(f(\mathcal{S}_1), f(\mathcal{S}_2)) \geq 1/2 f(\mathcal{S}^*) = 1/2 \, \mathrm{OPT} \, , \tag{4}$$

  so we have 1/2-approximation.

  Let $B = \{e_\ell : \ell \in \{1, 2, \ldots, n\}\}$ be the canonical basis in $\mathbb{R}^n$, where $n = |V|$. For every vertex $i \in V$ there is a vector x_i that is equal to e_k if node i is assigned to cluster k. Consider the following program:

$$\max \left\{ \sum_{\{i,j\} \in E} \left(w_{\{i,j\}}^+ x_i \cdot x_j + w_{\{i,j\}}^- (1 - x_i \cdot x_j) \right) : x_i \in B, \forall i \in V \right\} . \tag{5}$$

- Question 3. Explain why this program is a formulation of the correlation clustering problem.

  We have two cases. Either i and j belong to the same cluster, or either two different clusters. In the first case, there is some cluster k so that both i and j belong to k. We have $x_i = x_j = e_k$ and $x_i \cdot x_j = 1$. Also $e = \{i, j\} \in E(S_k)$. The edge e then contributes w_e^+ to the first term of both objective (1) and objective (5). In the second case we have $x_i \cdot x_j = 0$, since different basis vectors are orthogonal, as well as $e \in E \setminus \cup E(S)$ and the edge e contributes w_e^- to the second term of both objective (1) and objective (5). It follows that objective (1) and objective (5) are identical.

The formulation is relaxed to obtain the following vector program:

$$\max \left\{ \sum_{\{i,j\} \in E} \left(w^+_{\{i,j\}} v_i \cdot v_j + w^-_{\{i,j\}} (1 - v_i \cdot v_j) \right) \right\}. \tag{6}$$

subject to

$$\begin{aligned} v_i \cdot v_i &= 1\,, && \forall i \in V, \\ v_i \cdot v_j &\geq 0\,, && \forall i, j \in V, \\ v_i &\in \mathbb{R}^n\,, && \forall i \in V\,. \end{aligned} \tag{7}$$

Consider the following algorithm:

Algorithm SDP

Solve the the previous relaxation to obtain vectors $\{v_i : i \in V\}$, with objective value equal to Z. Draw independently two random hyperplanes with normals r_1 and r_2. This determines four regions,

$$\begin{aligned} R_1 &= \{i \in V : r_1 \cdot v_i \geq 0 \text{ and } r_2 \cdot v_i \geq 0\}, \\ R_2 &= \{i \in V : r_1 \cdot v_i \geq 0 \text{ and } r_2 \cdot v_i < 0\}, \\ R_3 &= \{i \in V : r_1 \cdot v_i < 0 \text{ and } r_2 \cdot v_i \geq 0\}, \\ R_4 &= \{i \in V : r_1 \cdot v_i < 0 \text{ and } r_2 \cdot v_i < 0\}, \end{aligned} \tag{8}$$

and output the partition $\mathcal{R} = \{R_1, R_2, R_3, R_4\}$.

In the following, the goal is to analyse this algorithm, and to prove that it is a 3/4-approximation.

- Question 4. Let $X_{\{i,j\}}$ be the random variable that is equal to 1 if the vectors v_i and v_j lie on the same side of the two random hyperplanes, and zero otherwise. Using an argument similar to the one used for Max-Cut, prove that $\text{Prob}(X_{\{i,j\}} = 1) = \left(1 - 1/\pi\theta_{\{i,j\}}\right)^2$, where $\theta_{\{i,j\}} = \arccos(v_i \cdot v_j)$ is the angle between vectors v_i and v_j.

  As in the lectures the change that one random hyperplane separates the two vectors is $\theta_{\{i,j\}}/\pi$, so the change that one hyperplane does not separate them is $1 - \theta_{\{i,j\}}/\pi$. Since the hyperplanes are chosen independently the change that neither of them separates them is

$$(1 - \theta_{\{i,j\}}/\pi)^2\,. \tag{9}$$

- Question 5. Let $f(\mathcal{R}) = \sum_{\{i,j\} \in E} \left(w^+_{\{i,j\}} X_{\{i,j\}} + w^-_{\{i,j\}} (1 - X_{\{i,j\}}) \right)$ the value of the partition $\mathcal{R}$, and denote $g(\theta) = (1 - \theta/\pi)^2$ the probability function computed before. Prove that the expected value of $f(\mathcal{R})$, denoted by $E(f(\mathcal{R}))$, is

$$\sum_{\{i,j\} \in E} w^+_{\{i,j\}} g(\theta_{\{i,j\}}) + w^-_{\{i,j\}} (1 - g(\theta_{\{i,j\}})). \tag{10}$$

  We have immediately

$$\begin{aligned} E(f(\mathcal{R})) &= \sum_{\{i,j\} \in E} \left(w^+_{\{i,j\}} E(X_{\{i,j\}}) + w^-_{\{i,j\}} (1 - E(X_{\{i,j\}})) \right) \\ &= \sum_{\{i,j\} \in E} w^+_{\{i,j\}} g(\theta_{\{i,j\}}) + w^-_{\{i,j\}} (1 - g(\theta_{\{i,j\}})). \end{aligned} \tag{11}$$

The following lemma will be helpful to conclude the analysis (You don't need to prove it.)

Lemma. For $\theta \in [0, \pi/2], g(\theta) \geq 3/4 \cos(\theta)$ and $1 - g(\theta) \geq 3/4(1 - \cos(\theta))$.

- Question 6. Using the lemma conclude that $E(f(\mathcal{R})) \geq 3/4\, Z$, and that the algorithm is a 3/4-approximation.

  Using Q5 and the lemma we find:

$$\begin{aligned} E(f(\mathcal{R})) &= \sum_{\{i,j\}\in E} w^+_{\{i,j\}} g(\theta_{\{i,j\}}) + w^-_{\{i,j\}}(1 - g(\theta_{\{i,j\}})) \\ &\geq 3/4 \sum_{\{i,j\}\in E} w^+_{\{i,j\}} \cos(\theta_{\{i,j\}}) + w^-_{\{i,j\}}(1 - \cos(\theta_{\{i,j\}})) = 3/4\, Z \geq 3/4\, \mathrm{OPT}\,, \end{aligned} \tag{12}$$

  so that

$$E(f(\mathcal{R})) \geq 3/4 \mathrm{OPT}\,. \tag{13}$$