1 Approximation Algorithm II Homework 1

2 Approximation Algorithm II Homework 2

3 Approximation Algorithm II Homework 3

4 Approximation Algorithm II Homework 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^- representing how similar and dissimilar are the nodes i and j, respectively. For $S \subset V$, 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 in order to maximize

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

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 $\mathcal{S}_1 = \{\{i\} : i \in V\}$, $\mathcal{S}_2 = \{V\}$ two extreme clusters (containing single or all). Output max of $f(\mathcal{S}_1)$ and $f(\mathcal{S}_2)$.

Problem 4.1 *Compute $f(\mathcal{S}_1)$ and $f(\mathcal{S}_2)$.*

Solution: Because no loop, hence no self loop. We have,

$$f(\mathcal{S}_1) = \sum_{e \in E} w_e^-$$

and

$$f(\mathcal{S}_2) = \sum_{e \in E} w_e^+$$

□

Problem 4.2 *Prove it's 2-approximation.*

Solution: It's obviouse

$$f(\mathcal{S}_{optimal}) \leq \sum_{e\in E} w_e^+ + \sum_{e\in E} w_e^- = f(\mathcal{S}_1)+f(\mathcal{S}_2) \leq 2\cdot\max(f(\mathcal{S}_1), f(\mathcal{S}_2))$$

Hence, the output is 2 optimal. □

Let $B = \{e_l : l \in \{1,\ldots,n\}\}$ be standard basis of $\mathbb{R}^n$. $n = |V|$. $\forall i \in V$ let $x_i = e_k$ if $i \in S_k \in \mathcal{S}$. Consider following:

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

Problem 4.3 *Explain why this program is a formulation of the correlation clustering problem.*

Solution: Because all x_i's are orthonormal, hence inner product are either 1 or 0, depending on if they are in the same cluster or not. Hence the objective function equal to the original objective function. □

We relax it to the following:

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

subject to

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

And **Algorithm SDP** Solve above to obtain v_i with objective value Z. Draw independently two random hyperplane with normals r_1 and r_2, determine 4 regions:

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

Output 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.

Problem 4.4 *Let $X_{i,j}$ be random variable, $X_{ij} = 1_{v_i,v_j \in R_1} + 1_{v_i,v_j \in R_4}$. Prove*

$$P[X_{ij} = 1] = (1 - \frac{1}{\pi}\theta_{ij})^2$$

where $\theta_{ij} = \arccos(v_i \cdot v_j)$.

Proof: As in the slides, for r_1, there are $(\pi - \theta_{ij})/\pi$ chance to be in the same side, and for r_2 also $(\pi - \theta_{ij})/\pi$ chance on the same side, hence

$$P[X_{ij} = 1] = (1 - \frac{1}{\pi}\theta_{ij})^2$$

□

Problem 4.5 *Let*

$$f(\mathcal{R}) = \sum_{\{i,j\}\in E} \left(w^+_{\{i,j\}} X_{\{i,j\}} + w^-_{\{i,j\}} (1 - X_{\{i,j\}}) \right)$$

denote by $g(\theta) = (1 - \frac{1}{\pi}\theta)^2$. Prove

$$E[f(\mathcal{R})] = \sum_{\{i,j\}\in E} \left(w^+_{\{i,j\}} g(\theta_{\{i,j\}}) + w^-_{\{i,j\}} (1 - g(\theta_{\{i,j\}})) \right)$$

.

Proof: Substitude Q4 to the formulas, using the linearity of expection, we have the result. □

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

Problem 4.6 *Prove $E[f(\mathcal{R})] \geq 3/4 \cdot Z$. And Z up-bounds the OPT, hence it's $3/4$-optimal.*

Solution: By setting of SDP, $v_i \cdot v_j \geq 0$, hence $\theta_{\{i,j\}} \in [0, \pi/2]$. Recall

$$Z = \sum_{\{i,j\}\in E} \left(w^+_{\{i,j\}} \cos(\theta_{ij}) + w^-_{\{i,j\}} (1 - \cos(\theta_{ij}) \right)$$

Hence

$$Z \leq \sum_{\{i,j\}\in E} \left(w_{\{i,j\}} 4/3 g(\theta_{ij}) + w^-_{\{i,j\}} 4/3(1 - g(\theta_{ij}) \right) = 4/3 E[f(\mathcal{R})]$$

□