CSCI-GA 3033-114/CS-GY 9223

Spring 2025

Homework 1

This homework is due **Thurs. Feb. 13 at 11:59 P.M.** You do not have to typeset your answers if you don't want to, but make sure they are legible.

Please do your homework together with one other person in the class. You and your partner should hand in only one copy of your solutions, through Gradescope, with both of your names on it. Use the group submission option when uploading.

POLICY ON CONSULTING REFERENCES: Please try to solve the problem yourselves without using any external references (you will learn a lot more that way). If you do consult external references, please write your solutions in your own words and cite any references you have used.

---

1. **Submodular Set Cover via LPs.** Let $I = \{1, \ldots, n\}$ be a set of n items, and let $u : 2^I \to \mathbb{Z}^{\geq 0}$ be a corresponding monotone submodular utility function, where $u(\emptyset) = 0$. Let $w_1, \ldots, w_n$ be non-negative weights associated with items in I.

   For $S \subseteq I$ and $i \in I$, let the marginal value of item i with respect to set S be $u_S(i) = u(S \cup \{i\}) - u(S)$. Let $Q = u(I)$ be the total value of set I.

   Consider the following IP, with variables $x_1, \ldots, x_n$:

$$\text{Minimize} \sum_{i \in I} w_i x_i \tag{1}$$

$$\sum_{i \in I \setminus S} u_S(i)\ x_i \geq Q - u(S) \qquad \forall S \subseteq I \tag{2}$$

$$x_i \in \{0, 1\} \qquad \forall i \in I. \tag{3}$$

   (Note that the number of constraints in this IP is exponential in n.)

   (a) Let $I' \subseteq I$ and let $\hat{x} \in \{0,1\}^n$ be the assignment to the variables x_i such that $\hat{x}_i = 1$ if $i \in I'$ and $\hat{x}_i = 0$ otherwise.

   Show that $I' \subseteq I$ is an optimal solution to the submoular cover problem for utility function u and item weights $w_1, \ldots, w_n$ iff $\hat{x}$ is an optimal solution to the above IP.

   (b) Give a counterexample to show that the statement in part (a) would not be true if we only included the constraint for $S = \emptyset$ in the IP. Note that this constraint is equivalent to $\sum_{i:x_i=1} u(\{i\}) \geq Q$.

(c) By changing the integer constraints $x_i \in \{0, 1\}$ to non-negativity constaints $x_i \geq 0$, we get the LP relaxation of the above IP. Call this relaxation LP1 and call its dual LP2. The variables of LP2 are y_S, for all $S \subseteq \emptyset$.

Write out the dual LP.

(d) A primal-dual algorithm for the submodular cover problem is as follows[1]:

- $I' = \emptyset$
- For all variables y_S in LP2, set $y_S = 0$
- While $u(I') \neq Q$
  - Increase variable $y_{I'}$ until some constraint of LP2 becomes tight. Let $i \in I$ be the item associated with that constraint.
  - $I' = I' \cup \{i\}$
- return I'

Prove that the above algorithm successfully terminates with a set I' such that $u(I') = Q$.

(e) The step of the algorithm in which $y_{I'}$ is increased can be viewed as a greedy step in which the "best" i is chosen. Rewrite this step as a minimization over all $i \in I \backslash S$ that takes time polynomial in n (assuming that a call to the value oracle for u takes constant time).

(f) Prove that the primal-dual algorithm above returns a cover I' such that

$$\frac{\sum_{i \in I'} w_i}{OPT} \leq \max_{S \subseteq I : u(S) \neq Q} \frac{\sum_{i \in I} u_S(i)}{Q - u(S)}.$$

That is, show that the above algorithm achieves an approximation factor of

$$\max_{S \subseteq I : u(S) \neq Q} \frac{\sum_{i \in I} u_S(i)}{Q - u(S)}.$$

(g) Show that if u is a coverage function corresponding to a set system with universe $\mathcal{U}$ and family of subsets $\mathcal{F}$, then $\max_{S \subseteq I} \frac{\sum_S u_S(i)}{Q - S(i)}$ is upper bounded by the value we've called f, the maximum number of sets in $\mathcal{F}$ that contain a single element in $\mathcal{U}$.

2. **(A Medical Problem.)** The input to this problem is a dataset derived from the medical records of m patients. It consists of m vectors $V = \{v_1, \ldots, v_m\}$, with each $v_i \in \{1, \ldots, d\}^n$. Each vector corresponds to the medical record of a single patient as follows. The record has n attributes, such as height, weight, and cholesterol. The value of each attribute has been discretized and converted into a value between 1 and d. In addition, each vector v_i in the dataset has a label $y_i \in \{+, -\}$, indicating whether or not the patient has a particular disease.

[1] Williamson and Shmoys define a primal-dual algorithm as follows (p. 24): "Primal-dual algorithms start with a dual feasible solution, and use dual information to infer a primal, possibly infeasible solution. If the primal solution is indeed infeasible, the dual solution is modified to increase the value of the dual objective function."

Let P be the set of vectors in the dataset that are labeled $+$, and let N be those labeled $-$. Assume that if two vectors in the dataset are identical, then they have the same label.

Define a subset S of the n attributes to be a *distinguishing set* for the dataset if the following holds: for all $v, v' \in V$, if v and v' have the same values on all attributes in S, then v and v' also have the same label. Intuitively, this means that for any $v \in V$, knowing just the values v has for the attributes in S (and not for any of the other attributes) is enough information to determine whether $v \in P$ or $v \in N$.

The problem is to find the smallest distinguishing set for such a dataset.

(a) Prove that this problem is NP-hard.

(b) Prove that this problem has a polynomial-time approximation algorithm with an approximation factor of $O(\log m)$.

3. **(Randomness Again.)** In Lecture #1 we saw a randomized rounding algorithm for set cover with expected cost $(1 + \ln n)\,OPT$. We now show how to do better when the sets have size at most $B \leq n$.

   - Solve the LP relaxation and get variables $x_S \in [0, 1]$.
   - Define $p_S := \min(1, x_S \ln B)$. Pick each set S independently with probability p_S.
   - For each element $e \in U$ uncovered by the random choices, pick the cheapest set containing it.

   Show that the expected cost of this algorithm is at most $O(\ln B) \cdot OPT$. For full points, show that the expected cost is at most $(1 + \ln B) \cdot LP$ (else you will get 80% of the points).

   (The greedy algorithm for weighted set cover, presented in Lecture #2, can also be shown to be an H_B-approximation, where $H_B \leq 1 + \ln B$ is the B^{th} Harmonic number. You can see a proof of the result in the Williamson-Shmoys book.)