AM 221: Advanced Optimization Spring 2016

Prof. Yaron Singer *Problem Set 9 — Due Wed, Apr. 6th at 23:59*

Instructions: All your solutions should be prepared in LaTeX and the PDF and .tex should be submitted to canvas. For each question, the best and correct answers will be selected as sample solutions for the entire class to enjoy. If you prefer that we do not use your solutions, please indicate this clearly on the first page of your assignment.

The programming parts can be written in the programming language of your choice and the code should be submitted alongside your solutions.

1. Properties of Submodular Functions.

a. Recall that for a finite set N, a function $f : 2^N \to \mathbb{R}$ is defined to be *submodular* if $\forall S, T \subseteq N$, such that $S \subseteq T$ and $\forall u \in N \setminus T$, the following holds:

$$f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T)$$

Prove that any non-negative submodular function is also subadditive, i.e. if $f : 2^N \to \mathbb{R}_{\geq 0}$ is submodular then $f(S \cup T) \leq f(S) + f(T)$ for any $S, T \subseteq N$.

b. Prove that a function $f : 2^N \to \mathbb{R}_{\geq 0}$ is submodular if and only if for any $S \subseteq N$ the marginal contribution function f_S defined by $f_S(T) = f(S \cup T) - f(S)$ (for $T \subseteq N$) is subadditive.

2. Approximate Oracles. Recall the greedy algorithm for maximizing a non-negative monotone submodular function described below.

Algorithm 1 Greedy Algorithm

1: Set $S = \emptyset$
2: **while** $|S| < k$ **do**
3: $\quad S \leftarrow S \cup \text{argmax}_{a \in N} f_S(a)$
4: **end while**
5: **return** S

For a given $\alpha < 1$ and set $S \subseteq N$ we will say that $\tilde{f}_S$ is an α-approximate marginal oracle if for any $T \subseteq N$ we have the guarantee that $\alpha f_S(T) \leq \tilde{f}_S(T) \leq f_S(T)$, for a given constant $\alpha < 1$. Prove that a greedy algorithm like the one above which uses an α-approximate marginal oracle in every iteration in line (3) provides a $1 - 1/e^{\alpha}$ approximation to the optimal solution.

3. Estimation of Marginal Influence using Chernoff Bounds. Recall that the algorithm for submodular maximization we used in class assumed that we can evaluate the marginal contribution

$f(S \cup v) - f(S)$ *exactly* for any $S \subseteq V$ and any $v \in V$. As we've seen in class, our representation of the Independent Cascade process as an influence maximization function f required summing over *exponentially* many coverage functions and therefore evaluating the marginal contributions *exactly* is computationally infeasible. In this exercise, we will see how to overcome this difficulty by using *sampling*: instead of computing the marginal contributions exactly, we will use sampling to *approximate* the marginal contributions, and bound the error on our evaluation of the marginal contribution calculations. We will then consider a modified greedy algorithm which uses approximations of the marginal contributions and show that the proof we showed in class for the approximation guarantee of the greedy algorithm for submodular maximization can easily be modified to the case where the algorithm has bounded errors in calculating the marginal contributions.

a. Consider the following algorithm for sampling the influence $f(v)$ of a single node v in the Independent Cascade model:

SampleInfluence
input: Graph $G = (V, E)$, edge probabilities $\{p_{u,w}\}_{(u,w)\in E}$, source node v, sample limit $m \in \mathbb{N}$ For $i \in \{1, 2, \ldots, m\}$: Realize every edge in $(u, w) \in E$ with probability $p_{u,w}$ and set E' to be the set of realized edges; Set r_i to be the number of nodes reachable (via, say, a BFS search) from v in $G' = (V, E')$ **return:** $\frac{1}{m}\sum_{i=1}^{m} r_i$

Let us denote by $\tilde{f}(v)$ the value returned by the algorithm. Find the number m of samples required to guarantee:

$$(1-\varepsilon)f(v) \leq \tilde{f}(v) \leq (1+\varepsilon)f(v)$$

with probability $1-\gamma$ for some $\varepsilon > 0$ and $\gamma > 0$ (m is a function of ε and γ). To do so, use the following version of the famous Chernoff bound:

Theorem 1. *Let $X_1, X_2, \ldots, X_m$ be independent random variables s.t. for every $i \in [m]$ we have that $X_i \in [0, b]$, and let $X = \sum_{i=1}^{m} X_i$ and $\mu = E[X]$ then for any $\delta \geq 0$:*

$$\mathbb{P}[|X - \mu| \geq \delta] \leq 2e^{\frac{-\delta^2}{mb^2}}$$

[**Hint:** the Chernoff bound gives an additive approximation whereas you are being asked to give a multiplicative approximation. For this, use a simple upper bound on $f(v)$.

b. Using ideas similar to part a., give an algorithm which given a set S and a node v computes an estimate $\tilde{f}_S(v)$ of the marginal contribution $f_S(v)$ such that:

$$(1-\varepsilon)f_S(v) \leq \tilde{f}_S(v) \leq (1+\varepsilon)f_S(v)$$

with probability $1-\gamma$ for some $\varepsilon > 0$ and $\gamma > 0$. What is the running time complexity of this algorithm as a function of ε, γ and the size of the graph.

c. Show that:

$$(1-\varepsilon)f_S(v) \leq \tilde{f}_S(v) \leq (1+\varepsilon)f_S(v)$$

Implies:

$$(1-2\varepsilon)g_S(v) \leq \tilde{f}_S(v) \leq g_S(v)$$

for g the function defined by $g(S) = (1+\varepsilon)f(S)$ (note that g is also submodular!).

d. Using parts b., c. and Problem 2, give a modified Greedy algorithm for maximizing influence in the Independent Cascade model which runs in time that is polynomial in ϵ, and the number of nodes in the graph and guarantees to find an approximation of $(1 - \frac{1}{e} - \epsilon)$ with probability $1 - \delta$, for any given $\epsilon > 0$ and $\delta > 0$. Prove the bound on the approximation ratio.

4. Influence Maximization in Social Networks. In this problem, we will consider a simplified model of influence in social networks: the social network is an undirected graph $G = (V, E)$ and for a set of nodes $S \subseteq V$, we denote by $N(S)$ the set of their neighbors:

$$N(S) = \{v \in V \mid \exists u \in S,\ (u, v) \in E\}$$

The influence $I(S)$ of a set of nodes is defined by $I(S) = |N(S)|$.

a. As the designer of a marketing campaign, your goal is to find a subset $S \subseteq V$ of at most K nodes whose influence is maximal. Give an approximation algorithm for this problem, what is its approximation ratio?

We will be using the dataset available at http://thibaut.horel.org/facebook.txt. This dataset is subgraph of the Facebook social graph. Each line in the file contains the id of two users, indicating that these two users are friends on Facebook.

b. Write a function which given the social network described in the dataset and a budget $K \in \mathbb{N}^+$ returns an approximately optimal set of nodes S for the influence function $I(S)$. The function should return both the users to influence and the value (total amount of influence) obtained.

c. Plot the influence $I(S)$ obtained by your function as a function of the budget K.