Lecturer: Hung Q. Ngo
Last update: November 14, 2006

Analyzing approximation algorithms with the dual-fitting method

1 A greedy algorithm for WEIGHTED SET COVER

One of the best examples of combinatorial approximation algorithms is a greedy algorithm approximating the WEIGHTED SET COVER problem. An instance of the SET COVER problem consists of a universe set U of m elements, a family $\mathcal{S}$ of n subsets of U, where set $S \in \mathcal{S}$ is weighted with w_S. We want to find a sub-family of $\mathcal{S}$ with minimum total weight such that the union of the sub-family is U (i.e. covers U). Consider the following greedy algorithm.

Algorithm 1.1. GREEDY-SET-COVER$(U, \mathcal{S}, w)$

1: $\mathcal{C} \leftarrow \emptyset, \quad A \leftarrow U$
2: **while** $A \neq \emptyset$ **do**
3: Pick $S \in \mathcal{S}$ with the least cost per un-covered element, i.e. pick S such that $\frac{w_S}{|S \cap A|}$ is minimized.
4: $A \leftarrow A - S$
5: $\mathcal{C} \leftarrow \mathcal{C} \cup \{S\}$
6: **end while**
7: **return** $\mathcal{C}$

In this section, we analyze this algorithm combinatorially. Then, a linear programming based analysis will be derived in the next section.

Without loss of generality, suppose the algorithm returns a collection $\{S_1, \dots, S_k\}$ of k sets. Let X_i be the set of newly covered elements of U after the ith step. Let $x_i = |X_i|$, and $w_i = w_{S_i}$ which is the weight of the ith set picked by the algorithm. Assign a cost $c(u) = w_i/x_i$ to each element $u \in X_i$, for all $i \leq k$.

For any set $S \in \mathcal{S}$, we first estimate $\sum_{u \in S} c(u)$. Let $a_i = |S \cap X_i|$. Then, it is easy to see the following:

$$\begin{aligned} \frac{w_S}{a_1 + \cdots + a_k} &\geq \frac{w_1}{x_1} \\ \frac{w_S}{a_2 + \cdots + a_k} &\geq \frac{w_2}{x_2} \\ \vdots \quad &\vdots \quad \vdots \\ \frac{w_S}{a_k} &\geq \frac{w_k}{x_k}. \end{aligned}$$

Hence,

$$\sum_{u \in S} c(u) = \sum_{i=1}^{k} a_i \frac{w_i}{x_i} \leq \sum_{i=1}^{k} a_i \frac{w_S}{a_i + \cdots + a_k} \leq w_S \cdot H_{|S|},$$

where $H_{|S|} = 1 + 1/2 + \cdots + 1/|S|$ is the $|S|$th harmonic number. Since $|S| \leq m$ for all S, we conclude that

$$\sum_{u \in S} c(u) \leq H_m \cdot w_S, \quad \forall S \in \mathcal{S}. \tag{1}$$

One may ask, what if $a_i + \cdots + a_k = 0$ for some i. This is not a problem. Since $S \neq \emptyset$, $a_1 + \cdots + a_k \neq 0$. If $a_i + \cdots + a_k = 0$ for some i, then all the terms $a_i \frac{w_i}{x_i}, \dots, a_k \frac{w_k}{x_k}$ can be ignored.

Let $\mathcal{T}$ be any optimal solution, then

$$\text{cost}(\mathcal{C}) = \sum_{u \in U} c(u) \leq \sum_{T \in \mathcal{T}} \sum_{u \in T} c(u) \leq \sum_{T \in \mathcal{T}} H_{|T|} \cdot w_T \leq H_m \cdot \text{cost}(\mathcal{T}).$$

We thus have proved the following theorem.

Theorem 1.2. GREEDY-SET-COVER *has approximation ratio* H_m.

Exercise 1. In the SET MULTICOVER problem, each element u is required to be covered m_u times, where m_u is a positive integer. Each set can be picked multiple times. The cost of picking S k times is kw_S. Devise a greedy algorithm for SET MULTICOVER with approximation ratio H_m (and prove that!).

Exercise 2. In the MAXIMUM COVERAGE problem, we are given a universe U, a collection $\mathcal{S}$ of subsets of U, and a positive integer k. Each element u in the universe has a non-negative integer weight w_u. The problem is to find k members of $\mathcal{S}$ whose union has the maximum total weight.

Suppose we solve this problem by greedily pick the best set in each iteration until k sets are picked. ("Best" set is the set maximizing total weight of uncovered elements.) Prove that this strategy has approximation ratio $1 - \left(1 - \frac{1}{k}\right)^k$.

Exercise 3. Consider the WEIGHTED VERTEX COVER problem in which each vertex v is weighted with $w_v > 0$. Consider the following algorithm

Algorithm 1.3. LR VERTEX COVER(G, w)

$C = \emptyset$
For each $v \in V(G)$, let $c(v) \leq w_v$
while C is not a vertex cover **do**
    Pick an uncovered edge (u, v), let $\epsilon \leq \min\{c(u), c(v)\}$
    $c(u) \leftarrow c(u) - \epsilon$; $c(v) \leftarrow c(v) - \epsilon$
    Add into C all vertices v having $c(v) = 0$.
end while
return C

Prove that this is a 2-approximation algorithm.

2 Analyzing GREEDY SET COVER with dual-fitting

It is natural to find out how Algorithm 1.1 relates to the integer programming formulation of SET COVER. The IP for SET COVER is

$$\begin{array}{rll} \min & \displaystyle\sum_{S \in \mathcal{S}} w_S x_S & \\ \text{subject to} & \displaystyle\sum_{S \ni u} x_S \geq 1, & \forall u \in U, \\ & x_S \in \{0, 1\}, & \forall S \in \mathcal{S}. \end{array} \tag{2}$$

The LP-relaxation is

$$\begin{array}{rll} \min & \displaystyle\sum_{S \in \mathcal{S}} w_S x_S & \\ \text{subject to} & \displaystyle\sum_{S \ni u} x_S \geq 1, & \forall u \in U, \\ & x_S \geq 0, & \forall S \in \mathcal{S}. \end{array} \tag{3}$$

And, the dual LP is

$$\begin{array}{lll} \max & \sum_{u \in U} y_u & \\ \text{subject to} & \sum_{u \in S} y_u \le w_S, & \forall S \in \mathcal{S}, \\ & y_u \ge 0, & \forall u \in U. \end{array} \tag{4}$$

The dual constraints look very much like relation (1), except that we need to divide both sides of (1) by H_m. Thus, for each $u \in U$, if we set $y_u = c(u)/H_m$, then $\mathbf{y}$ is a dual feasible solution. It follows that

$$\text{cost}(\mathcal{C}) = \sum_{u \in U} c(u) = H_m \, \text{cost}(\mathbf{y}) \le H_m \cdot \text{OPT}.$$

3 More general covering problems

The CONSTRAINED SET MULTICOVER problem is a generalization of the SET COVER problem in which each elements $u \in U$ needs to be covered m_u times, where m_u is a positive integer.

The corresponding integer program can be written as

$$\begin{array}{lll} \min & \sum_{S \in \mathcal{S}} w_S x_S & \\ \text{subject to} & \sum_{S \ni u} x_S \ge m_u, & \forall u \in U, \\ & x_S \in \{0, 1\}, & \forall S \in \mathcal{S}. \end{array} \tag{5}$$

When relaxing this program, it is no longer possible to remove the upper bounds $x_S \le 1$ (otherwise an integral optimal solution to the LP may not be an optimal solution to the IP). The LP-relaxation is

$$\begin{array}{lll} \min & \sum_{S \in \mathcal{S}} w_S x_S & \\ \text{subject to} & \sum_{S \ni u} x_S \ge m_u, & \forall u \in U, \\ & -x_S \ge -1, & \forall S \in \mathcal{S}, \\ & x_S \ge 0, & \forall S \in \mathcal{S}. \end{array} \tag{6}$$

The dual linear program is now

$$\begin{array}{lll} \max & \sum_{u \in U} m_u y_u - \sum_{S \in \mathcal{S}} z_S & \\ \text{subject to} & \sum_{u \in S} y_u - z_S \le w_S, & \forall S \in \mathcal{S}, \\ & y_u, z_S \ge 0, & \forall u \in U, \forall S \in \mathcal{S}. \end{array} \tag{7}$$

We will try to devise a greedy algorithm to solve this problem and analyze it using the dual-fitting method.

Algorithm 3.1. GREEDY-SET-MULTICOVER$(U, \mathcal{S}, w, m)$

1: $\mathcal{C} = \emptyset$; $A \leftarrow U$
2: // We call an element $u \in U$ "alive" if $m_u > 0$. Initially all of A are alive
3: **while** $A \ne \emptyset$ **do**
4: Pick S such that $\frac{w_S}{|S \cap A|}$ is minimized.
5: $\mathcal{C} = \mathcal{C} \cup \{S\}$

    $m_u \leftarrow m_u - 1$ for each $u \in S \cap A$
    Remove from A all elements u with $m_u = 0$
end while
return $\mathcal{C}$

The next step is to write the cost of $\mathcal{C}$ as a multiple of the cost of a feasible solution to (7), namely cost$(\mathcal{C}) = \rho(\sum_u m_u y_u - \sum_S z_S)$ for some feasible solution $(\mathbf{y}, \mathbf{z})$ of (7). For each element $u \in U$ and each $j \in [m_u]$, let $c(u, j)$ be the cost of covering u for the jth time. If S covers u for the jth time, and A_S is the set of alive elements before S was picked, then $c(u, j) = \frac{w_S}{|S \cap A_S|}$. If S was chosen before T, then $A_T \subseteq A_S$. Thus,

$$\frac{w_S}{|S \cap A_S|} \leq \frac{w_T}{|T \cap A_S|} \leq \frac{w_T}{|T \cap A_T|}.$$

Consequently, for any u we have $c(u, 1) \leq \cdots \leq c(u, m_u)$. The final cost is

$$\text{cost}(\mathcal{C}) = \sum_{u \in U} \sum_{j=1}^{m_u} c(u, j).$$

In order to write this sum as $\rho(\sum_{u \in U} m_u y_u - \sum_{S \in \mathcal{S}} z_S)$ (keeping in mind that $y_u, z_S \geq 0$), it makes sense to try

$$\begin{aligned} \text{cost}(\mathcal{C}) &= \sum_{u \in U} m_u c(u, m_u) - \sum_{u \in U} \sum_{j=1}^{m_u - 1} [c(u, m_u) - c(u, j)] \\ &= \sum_{u \in U} m_u c(u, m_u) - \sum_{u \in U} \sum_{j=1}^{m_u} [c(u, m_u) - c(u, j)] \end{aligned}$$

The double sum (after the minus sign) is non-negative, which is good. We need to write it in the form $\rho \sum_{S \in \mathcal{S}} z_S$ somehow. Note that, each time u is covered, a term $c(u, m_u) - c(u, j)$ is added into the double sum. For each $S \in \mathcal{C}$, suppose S covers $u \in S \cap A_S$ the $j_{u,S}$th time. Then,

$$\sum_{u \in U} \sum_{j=1}^{m_u} [c(u, m_u) - c(u, j)] = \sum_{S \in \mathcal{C}} \sum_{u \in S \cap A_S} [c(u, m_u) - c(u, j_{u,S})].$$

Now, let ρ be a number to be determined. Define

$$\begin{aligned} y_u &= \frac{1}{\rho} c(u, m_u), \quad \forall u \in U \\ z_S &= \begin{cases} \frac{1}{\rho} \sum_{u \in S \cap A_S} [c(u, m_u) - c(u, j_{u,S})] & S \in \mathcal{C} \\ 0 & S \notin \mathcal{C} \end{cases} \end{aligned}$$

For $(\mathbf{y}, \mathbf{z})$ to be dual-feasible, we would like to find ρ so that, for each $S \in \mathcal{S}$, $\sum_{u \in S} y_u - z_S \leq w_S$.

Consider any $S \notin \mathcal{C}$. In this case,

$$\sum_{u \in S} y_u - z_S = \frac{1}{\rho} \sum_{u \in S} c(u, m_u).$$

Let $u_1, \ldots, u_k$ be the elements of S. Without loss of generality, assume that u_1 was completely covered before u_2, and so on. Then, right before u_i is completely covered, $|A_S| \geq k - (i - 1)$. Hence, $c(u_i, m_{u_i}) \leq w_S/(k - i + 1)$. Consequently,

$$\sum_{u \in S} y_u - z_S \leq \frac{1}{\rho} \sum_{i=1}^{k} \frac{w_S}{k - i + 1} \leq \frac{H_m}{\rho} \cdot w_S.$$

Now, consider any $S \in \mathcal{C}$. In this case we have

$$\begin{aligned}\sum_{u \in S} y_u - z_S &= \frac{1}{\rho}\sum_{u\in S} c(u, m_u) - \frac{1}{\rho}\sum_{u \in S \cap A_S} [c(u, m_u) - c(u, j_{u,S})] \\ &= \frac{1}{\rho}\left(\sum_{u \in S \setminus A_S} c(u, m_u) + \sum_{u \in S\cap A_S} c(u, j_{u,S})\right)\end{aligned}$$

Let $u_1, \dots, u_{k'}$ be elements in $S \setminus A_S$ which were completely covered in that order. Note that $0 \le k' < k$. Note also that $\sum_{u \in S \cap A_S} c(u, j_{u,S}) = w_S$. Similar to the previous reasoning, we get

$$\sum_{u \in S} y_u - z_S = \frac{1}{\rho}\left(\sum_{i=1}^{k'} \frac{w_S}{k-i+1} + w_S\right) \le \frac{H_m}{\rho} \cdot w_S.$$

Hence, $(\mathbf{y}, \mathbf{z})$ would be a dual feasible solution if we pick $\rho = H_m$, which is also an approximation ratio for Algorithm 3.1.

Exercise 4. Devise a greedy algorithm for SET MULTICOVER with approximation ratio H_m. Analyze your algorithm using the dual-fitting method.

Exercise 5. In the MULTISET MULTICOVER problem, we are given a collection $\mathcal{S}$ of multisets of a universe U. For each $S \in \mathcal{S}$, let $M(S, u)$ be the multiplicity of u in S. Each element u needs to be covered m_u times. We can assume $M(S, u) \le m_u$ for all S, u.

Devise a greedy algorithm for MULTISET MULTICOVER with approximation ratio H_d, where d is the largest multiset size. The size of a multiset is the total multiplicity of its elements. Analyze your algorithm using the dual-fitting method.

Exercise 6. Consider the integer program $\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \ge \mathbf{b}\}$, where $\mathbf{A}, \mathbf{b}$ have non-negative integral entries, and $\mathbf{x}$ is required to be non-negative and integral also. This is called a covering integer program.

Use scaling and rounding to reduce covering integer programs to MULTISET MULTICOVER, so that we can use the greedy algorithm for the MULTISET MULTICOVER instance to get a greedy algorithm for the COVERING INTEGER PROGRAM instance with approximation ratio $O(\lg n)$, where n is the input size of the covering integer program. (Thus, the instance of MULTISET MULTICOVER must have size polynomial in n.)

Exercise 7. Vazirani's book. Problem 24.12, page 241.

Historical Notes

The greedy approximation algorithm for SET COVER is due to Johnson [5], Lovász [6], and Chvátal [2]. Feige [4] showed that approximating SET COVER to an asymptotically better ratio than $\ln m$ is $\mathbf{NP}$-hard.

The dual-fitting analysis for GREEDY SET COVER was given by Lovász [6]. Dobson [3] and Rajagopalan and Vazirani [8] studied approximation algorithms for covering integer programs. The dual-fitting method has found applications in other places [1, 7].

References

[1] P. CARMI, T. ERLEBACH, AND Y. OKAMOTO, *Greedy edge-disjoint paths in complete graphs*, in Graph-theoretic concepts in computer science, vol. 2880 of Lecture Notes in Comput. Sci., Springer, Berlin, 2003, pp. 143–155.

[2] V. CHVÁTAL, *A greedy heuristic for the set-covering problem*, Math. Oper. Res., 4 (1979), pp. 233–235.

[3] G. DOBSON, *Worst-case analysis of greedy heuristics for integer programming with nonnegative data*, Math. Oper. Res., 7 (1982), pp. 515–531.

[4] U. FEIGE, *A threshold of* $\ln n$ *for approximating set cover (preliminary version)*, in Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996), New York, 1996, ACM, pp. 314–318.

[5] D. S. JOHNSON, *Approximation algorithms for combinatorial problems*, J. Comput. System Sci., 9 (1974), pp. 256–278. Fifth Annual ACM Symposium on the Theory of Computing (Austin, Tex., 1973).

[6] L. LOVÁSZ, *On the ratio of optimal integral and fractional covers*, Discrete Math., 13 (1975), pp. 383–390.

[7] M. MAHDIAN, E. MARKAKIS, A. SABERI, AND V. VAZIRANI, *A greedy facility location algorithm analyzed using dual fitting*, in Approximation, randomization, and combinatorial optimization (Berkeley, CA, 2001), vol. 2129 of Lecture Notes in Comput. Sci., Springer, Berlin, 2001, pp. 127–137.

[8] S. RAJAGOPALAN AND V. V. VAZIRANI, *Primal-dual RNC approximation algorithms for set cover and covering integer programs*, SIAM J. Comput., 28 (1999), pp. 525–540 (electronic). A preliminary version appeared in FOCS'93.