A Note on the Greedy Approximation Algorithm for the Unweighted Set Covering Problem

Abhay K. Parekh

Laboratory For Information and Decision Systems, M.I.T., Cambridge, MA 02139, USA

A simple greedy approximation algorithm for the unweighted set covering problem has been analyzed extensively in the literature. The common conclusion has been that in the worst case, the heuristic yields a set cover of size $K_o \sum_{i=1}^{d} \frac{1}{i}$, where d is the cardinality of the largest covering set and K_o is the optimal cover size. This bound is attained when $N = z\,z!$ and $K_o = z!$, where N is the cardinality of the set being covered and z is integer. We present here a bound that is is tight for all values of N and K_o. An interesting aspect of this bound is that it is tight for some special cases of the set covering problem as well. For example, for the dominating set problem, the bound is attained for all N and K_o, $N > K_o^{K_o+1}$, where N is the number of nodes in the graph and K_o is the domination number.
Procedures to construct instances for which the heuristic exhibits worst-case behaviour for the unweighted set covering and dominating set problems are also presented.

Keywords: **Set Covering Problem, Approximation Algorithms.**
AMS Subject Classification: **68 Computer Science, 05 Combinatorics.**

The unweighted set covering problem is the following: Given finite sets, S and F where $F = \{F_1, F_2, \ldots, F_p\}$, and $\cup_{i=1}^{p} F_i = S$, find a minimum cardinality subset f of F such that $\cup_{F_i \in f} F_i = S$. We denote $N = |S|$. Consider the following heuristic denoted *Greedy*: Define an element of S to be "covered" in the beginning of an iteration, if it is contained in at least one of the sets picked by the heuristic so far. Initially, no elements are covered. In each iteration, put into the set cover, the least indexed element of F that covers the maximum number of uncovered elements of S, until all such elements are covered. (Selecting the least numbered element is just a way of breaking ties.) This heuristic has been analyzed in the literature, and it has been shown that the worst-case fractional error is $\sum_{i=1}^{d} \frac{1}{i}$ where d is the maximum number of elements in any of the covering sets [1],[2],[3],[4]. This bound is attained when $N = zz!$, $K_o = z!$, for integer z. We present a bound that is attained for all values of N and K_o. The analysis includes an algorithm for generating worst–case instances for all values N and K_o.

Let the set picked by *Greedy* in the i^{th} iteration be s_i, and let there be d^* iterations. Define m_i to be the number of uncovered elements of S covered by s_i when it is picked by *Greedy*.

The following theorem establishes a convenient relationship between K_o and d^*.

Theorem 1. *If Greedy returns a set cover $S^* = \{s_1 \ldots s_{d^*}\}$ then:*

(a) $\sum_{i=1}^{d^*} m_i = N$;

(b) $m_1 \geq m_2 \geq ...m_{d^*} \geq 1;$

(c) $\sum_{i=1}^{p} m_i + K_o m_{p+1} \geq N \qquad p = 0, 1, ...d^* - 1.$

Proof: (a) and (b) follow directly from the definition of *Greedy*. At iteration $p+1$ there are exactly $N - \sum_{i=1}^{p} m_i$ uncovered elements. Let this set be U_{p+1}. By choice of s_{p+1}, no set covers more than m_{p+1} members of U_{p+1}. Now consider any optimal set cover, S^*, and let α be the member of this set that covers the maximum number of elements in U_{p+1} of all the sets in S^*. This number is no greater than m_{p+1}, but is certainly at least as great as the the average number of nodes in U_{p+1} which are covered by nodes in S^*.

$$\frac{|U_{p+1}|}{K_o} = \frac{N - \sum_{i=1}^{p} m_i}{K_o} \leq m_{p+1}, \quad p = 0, 1, ..., d^* - 1.$$

The result follows directly.

Now let T_z be a lower bound on the minimum number of elements of S which could ever be covered after $z \leq d^*$ iterations of *Greedy*. T_z is obtained by solving the following integer linear program, ILP:

$$T_z = \min \sum_{i=1}^{z} m_i \tag{1}$$

$$s.t$$

$$m_1 \geq m_2 \geq ... \geq m_z \geq 1, \tag{2}$$

$$\sum_{i=1}^{p} m_i + K_o m_{p+1} \geq N, \; p = 0, 1, ..., z-1. \tag{3}$$

$$m_i \in \text{Integers } i = 0, 1, ..., z. \tag{4}$$

Lemma 1. *Let an optimal solution be $q_1, q_2, \ldots, q_z$, and let $\exists j$ the largest integer $\leq z$ such that:*

$$q_j = \lceil \frac{N - \sum_{i=1}^{j-1} q_i}{K_o} \rceil + \Delta, \quad \Delta \geq 1.$$

Then the following solution is also optimal.

$$n_i = \begin{cases} q_i \,, & \text{if } i = 1, \ldots, j-1; \\ \lceil \frac{N - \sum_{p=1}^{i-1} n_p}{K_o} \rceil \,, & \text{if } i = j, \ldots, z; \end{cases}$$

Proof: We show that $\sum_{i=j}^{z} q_i \geq \sum_{i=j}^{z} n_i$ implying that $\sum_{i=1}^{z} q_i \geq \sum_{i=1}^{z} n_i$, i.e., the set of n_i is also optimal. Our approach is to proceed by induction on $\tau = z - j$.

$\underline{\tau = 0:}$ Observe that $q_z = n_z + \Delta$, implying that $\sum_{i=j}^{z} q_i \geq \sum_{i=j}^{z} n_i$.

$\underline{\tau = k+1:}$ We want to show that $\sum_{i=j}^{j+k} (q_i - n_i) \geq n_{j+k+1} - q_{j+k+1}$.

Let

$$\sum_{i=j+1}^{j+k} q_i = \alpha K_o + \beta, \qquad \sum_{i=j+1}^{j+k} n_i = \hat{\alpha} K_o + \hat{\beta}, \qquad \text{and} N - \sum_{p=1}^{j} n_p = \alpha^* K_o + \beta^*,$$

where $\alpha, \hat{\alpha}, \alpha^* \geq 0$ and $0 \leq \beta, \hat{\beta}, \beta^* < K_o$. Substituting we have:

$$(\alpha - \hat{\alpha})(K_o - 1) + \beta - \hat{\beta} + \Delta \geq \lceil \frac{\beta^* - \hat{\beta}}{K_o} \rceil - \lceil \frac{\beta^* - \beta - \Delta}{K_o} \rceil. \tag{5}$$

Consider the case when $\beta < \hat{\beta}$. Let $\gamma = \hat{\beta} - \beta, \quad 1 \leq \gamma \leq K_o - 1$. Substituting in (5) we must now show that:

$$(\alpha - \hat{\alpha})(K_o - 1) - \gamma + \Delta \geq \lceil \frac{\beta^* - \beta - \gamma}{K_o} \rceil - \lceil \frac{\beta^* - \beta - \Delta}{K_o} \rceil.$$

Observe that by the induction hypothesis, $\alpha - \hat{\alpha} \geq 1$. Thus:

$$K_o - 1 - \gamma + \Delta \geq \Delta \geq \lceil \frac{\beta^* - \beta - \gamma}{K_o} \rceil - \lceil \frac{\beta^* - \beta - \gamma - K_o \Delta}{K_o} \rceil \geq \lceil \frac{\beta^* - \beta - \gamma}{K_o} \rceil - \lceil \frac{\beta^* - \beta - \Delta}{K_o} \rceil.$$

Now suppose $\beta \geq \hat{\beta}$. Let $\gamma = \beta - \hat{\beta}$. Again substituting in (5) we have:

$$(\alpha - \hat{\alpha})(K_o - 1) + \gamma + \Delta \geq \lceil \frac{\beta^* - \hat{\beta}}{K_o} \rceil - \lceil \frac{\beta^* - \hat{\beta} - \Delta - \gamma}{K_o} \rceil.$$

But this is true because:

$$(\alpha - \hat{\alpha})(K_o - 1) + \gamma + \Delta \geq \gamma + \Delta \geq \lceil \frac{\beta^* - \hat{\beta}}{K_o} \rceil - \lceil \frac{\beta^* - \hat{\beta} - K_o \gamma - K_o \Delta}{K_o} \rceil \geq \lceil \frac{\beta^* - \hat{\beta}}{K_o} \rceil - \lceil \frac{\beta^* - \hat{\beta} - \Delta - \gamma}{K_o} \rceil. \blacksquare$$

Lemma 2. *An optimal set of m_i's is:*

$$m_i = \lceil \frac{N - \sum_{p=1}^{i-1} m_p}{K_o} \rceil \quad i = 1, 2, \ldots z. \tag{6}$$

Proof: By contradiction. Suppose this choice is not optimal for some N, K_o. Let $q_1, \ldots, q_z$ be an optimal solution, and let $q_{i_1}, \ldots, q_{i_I}$ be the set such that

$$q_{i_j} > \lceil \frac{N - \sum_{p=1}^{i_j - 1} q_p}{K_o} \rceil \qquad j = 1, 2, \ldots, I$$

We can now apply Lemma 1 I times to construct an optimal solution that is identical to the m_i's. But this contradicts our assumption, thus proving the lemma.

Combining this result with part (c) of Theorem 1 we have the result:

Theorem 2. *For any set covering problem:*

$$\frac{d^*}{K_o} \leq \frac{1}{K_o}(z : T_z = N) \tag{7}$$

Next, we present an algorithm for constructing instances of the set covering problem for which, given N and K_o, the bound in (7) holds. These instances have the pleasing property that if *Greedy* is run on them, the number of uncovered elements covered by the set picked in iteration i, is exactly m_i, i.e. obtained from (6). The approach is to create two partitions of S, one consisting of K_o sets and the other of d^* sets. We choose these sets so that *Greedy* picks the d^* sets even though the optimal set cover size is K_o.

Theorem 3. *The bound of Theorem 2 is attained for all values of N and K_o, $K_o \leq N$.*

Proof: The construction proceeds as follows:

[i] Let $S = \{1, 2, \ldots N\}$. Partition the elements of S into sets $G_0, G_1, \ldots G_{K_o-1}$ such that: $|G_i| = \lceil \frac{N}{K_o} \rceil \quad i = 0, 1, 2, \ldots, (N \bmod K_o) - 1$, and $|G_i| = \lfloor \frac{N}{K_o} \rfloor \quad i = (N \bmod K_o), \ldots, K_o - 1$

[ii] Define the sets $F_1, F_2, \ldots F_{d^*}$ and initialize them to be null sets.

[iii] Partition the elements of S into these sets by executing the following simple procedure:

Poll-The-G's
begin
p:=0;
for i:= 1 to d^* **do**
for j:= 1 to m_i (* The m_i's are obtained from (6) *)
begin
$F_i = F_i \cup \{\alpha\}, \quad \alpha \in G_p \cap \overline{(F_i \cap F_{i-1} \cap \ldots \cap F_1)}$
$p := (p+1) \bmod K_o;$
end
end

At the end of the procedure we have the sets $F_1, F_2, \ldots F_{d^*}$ such that $|F_i| = m_i, \quad i = 1, 2, \ldots d^*$ Since they fully partition the elements of S, the F_i's form a set cover of size d^*. It is easy to show by induction on i that

$$\max_j |G_j - \cup_{k=1}^{i} F_k| = |F_{i+1}|, \quad i = 1, 2, \ldots, d^* - 1 \tag{8}.$$

[iv] Let $F = \{F_1, \ldots F_{d^*+K_o}\}$ where $F_1, \ldots F_{d^*}$ are as defined above and $F_{d^*+\delta} = G_{\delta-1}$ for $\delta = 1, 2, \ldots, K_o$.

Now suppose *Greedy* is run on the constructed instance of the set covering problem. By definition, the minimum cover, f, is $\{F_{d^*+1}, F_{d^*+2}, \ldots F_{d^*+K_o}\}$, i.e., $|f| = K_o$. Since the heuristic picks the lowest indexed member of F of maximum cardinality, $s_1 = F_1$. At the end of the iteration we see from (8) that the maximum number of uncovered nodes in any element of f is just F_2, implying that $s_2 = F_2$. This continues until *Greedy* has picked $F_1, F_2, \ldots F_{d^*}$. Since S is partitioned over these sets, the heuristic terminates and we have met exactly the bound of Theorem 2.

Done.

Corollary 3.1. *If $N = zz!$, and $K_o = z!$, for some integer z, then $f^*(N, K_o) = \sum_{i=1}^{z} \frac{1}{i}$.*

Proof: This can be seen by substitution into (6). We see that $m_i = z$ for $i = 1 \ldots \frac{z!}{z}$; $m_i = z - 1$ for the next $\frac{z!}{z-1}$ values of i etc. For the last $z!$ values of i, $m_i = 1$. Notice that:

$$\sum_{i=1}^{d^*} m_i = \sum_{p=0}^{z-1} (z-p)\frac{z!}{z-p} = zz!$$

Thus, $d^* = z! \sum_{i=1}^{z} \frac{1}{i}$. The result follows from Theorem 3.
The bound of Theorem 2 exactly characterizes the worst-case performance of *Greedy*, but we do not have a closed-form expression for it. In our next few results we bound the worst-case performance of Greedy (denoted by $f^*(N, K_o)$) from above and below:

Theorem 4. *For any set covering problem:*

$$f^*(N, K_o) \leq K_o + \log_{\frac{K_o}{K_o - 1}} \left(\frac{N}{K_o} \right). \tag{9}$$

Proof: First, we claim that

$$m_i = \max \left\{ \frac{N - \sum_{i=1}^{i-1} m_i}{K_o}, 1 \right\} \qquad i = 1, ..., z \tag{10}$$

is an optimal solution to the integer relaxation of ILP. This can be seen by looking at the dual of the problem and applying complementary slackness conditions. The interested reader is encouraged to work out the details.
After some algebra we have:

$$m_i = \max \left\{ \frac{N}{K_o} \left(1 - \frac{1}{K_o} \right)^{i-1}, 1 \right\} \qquad i = 1, 2, \ldots, z.$$

For $z = d^*$, simplification yields that $m_i = 1, \forall i \geq i_{min}$ where

$$i_{min} = 1 + \log_{\frac{K_o}{K_o - 1}} \frac{N}{K_o}. \tag{11}$$

By summing the geometric series:

$$\sum_{j=1}^{i_{min}-1} m_j = N - K_o.$$

But we want to find $\hat{d} : T_{\hat{d}} = N$, since this will yield an upper bound on $f^*(N, K_o)$ (from Theorem 3). So, we have:

$$f^*(N, K_o) \leq \hat{d} = K_o + \log_{\frac{K_o}{K_o - 1}} \frac{N}{K_o}.$$

Done

Theorem 5. *Let $f^*(N, K_o)$ be the largest set cover returned by Greedy over all instances. Then for $K_o \geq 2$:*

$$f^*(N, K_o) > \log_{\frac{K_o}{K_o-1}} \frac{N}{K_o}.$$

Proof: Define the following:

$$m_i = \bar{m}_i + \Delta_i \qquad i = 1, 2, ..., d^*.$$

$$\bar{m}_i = \frac{N}{K_o}\left(1 - \frac{1}{K_o}\right)^{i-1} \qquad i = 1, 2, ..., d^* - 1.$$

$$m_i = \lceil \frac{N - \sum_{p=1}^{i-1} m_p}{K_o} \rceil i = 1, 2, ..., d^*.$$

The $\bar{m}_i$'s correspond to the values of the decision variables of the relaxation of ILP. The Δ_i's are the "error terms" associated with approximating the solution of ILP by its relaxation. First, observe that: $i = 1$: $m_1 \leq \frac{N}{K_o} + 1$ and $\bar{m}_1 = \frac{N}{K_o}$. So $\Delta_1 \leq 1$.

$$m_{p+1} = \lceil \frac{N - \sum_{j=1}^{p} m_j}{K_o} \rceil = \lceil \frac{N - \sum_{j=1}^{p} \bar{m}_j - \sum_{j=1}^{p} \Delta_j}{K_o} \rceil = \lceil \bar{m}_{p+1} - \frac{\sum_{j=1}^{p} \Delta_j}{K_o} \rceil$$

So we have $m_{p+1} \leq \bar{m}_{p+1} - \frac{\sum_{j=1}^{p} \Delta_i}{K_o} + 1$. i.e.,

$$\Delta_{p+1} \leq 1 - \frac{\sum_{j=1}^{p} \Delta_i}{K_o}$$

$$\sum_{j=i}^{p+1} \Delta_j \leq 1 + \sum_{j=1}^{p} \Delta_j \left(1 - \frac{1}{K_o}\right)$$

Solving the recurrence in terms of Δ_1, and setting $\Delta_1 = 1$:

$$\sum_{j=1}^{p} \Delta_j \leq \sum_{j=1}^{p} \left(1 - \frac{1}{K_o}\right)^{j-1}$$

The limit of this sum is K_o. Thus, $\sum_{i=1}^{d^*} m_i - \sum_{i=1}^{d^*} \bar{m}_i < K_o$. The result follows from Theorem 3.

Finally, we show that all our results apply to two special cases of the Set Covering Problem—the Directed and Undirected Dominating Set Problems. Here we are given a directed (undirected) graph, $G(V, A)$ with $V = \{1, 2, \ldots, N\}$, and we are to find the minimum cardinality set of nodes such that for every node, α that is not in the set there is at least one node in the set from which

there is an edge to α. The size of smallest dominating set is called the domination number. The following approximation algorithm, $GREEDY_{dom}$ is considered:
Define a node α to be "covered" in the beginning of an iteration, if at least one of the nodes picked by the heuristic so far has an edge to α. Initially, no elements are covered. In each iteration, put into the dominating set, the least numbered node that covers the maximum number of uncovered nodes, until all nodes are covered.
Let $S = \{1,2,\ldots,N\}$, $F_i = \{i\} \cup \{j : (i,j) \in A\}$ $i = 1,\ldots,N$, and let K_o be the domination number of the graph. Thus, we have an instance of the set covering problem to which *Greedy* can be applied. The resulting set cover can then be transformed to the dominating set that would be picked by $GREEDY_{dom}$. The bounds in Theorems 1–4 clearly apply, and we now show that the bound in Theorem 2 is attained, for values of $K_o \neq 2$ and $N \geq K_o^{K_o+1}$. For $K_o = 2$, the bound is attained for $N \geq 16$. Before proceeding, we give a simple Lemma that will be useful later:

Lemma 4. *If* $N \geq K_o^{K_o+1}$, $m_{K_o} \geq 2\ K_o$ *for* $K_o \neq 2$. *If* $K_o = 2$ *then* $m_2 \geq 4$ *for* $N \geq 16$.

Proof: The proof is simply by substitution into (6).

In what follows we consider the undirected version of the Dominating Set Problem. This is because given any undirected graph, we can convert it to a directed one by replacing every edge with two directed ones. The construction procedure used in the proof of Theorem 3 has to be modified to ensure that the F_i's correspond to the closed neighborhoods of the nodes of the graph i.e. $i \in N(j)$ iff $j \in N(i)$.

[i] Let $S = \{1,2,\ldots N\}$. ($N \geq 16$ if $K_o = 2$; else $N \geq K_o^{K_o+1}$.) Partition the elements of S into sets $G_0, G_1, \ldots G_{K_o-1}$ such that: $|G_i| = \lceil \frac{N}{K_o} \rceil$ $i = 0,1,2,\ldots,(N \bmod K_o) - 1$, and $|G_i| = \lfloor \frac{N}{K_o} \rfloor$ $i = (N \bmod K_o),\ldots,K_o - 1$

[ii] Pick $V_1 = \{v_1, v_2, \ldots v_{K_o}\}$ such that $v_i \in G_{i+1}$. (This will be the optimal dominating set.)

[iii] Define the sets $F_1, F_2, \ldots F_{d^*}$ and intialize them to be null sets.

[iv] We partition the elements of S into these sets by executing the following simple procedure:

Poll-The-G's-Carefully
begin
p:=0;
for i:= 1 to d^* **do**
begin
Picked-From-Opt = false;
for j:= 1 to m_i (* The m_i's are obtained from (6) *)
begin
if ($p = i - 1$) **AND** (**NOT** Picked-From-Opt) **then**
begin
$F_i = F_i \cup v_i$
Picked-From-Opt = true
end
else $F_i = F_i \cup \{\alpha\}$, $\alpha \in G_p - \{v_p\} \cap \overline{(F_i \cap F_{i-1} \cap \ldots \cap F_1)}$;
$p := (p+1) \bmod K_o$;
end

end

Note that this is a special case of the earlier construction and so the set of F_i's forms a set cover of $S = \{1, .., N\}$, as does the set of G_i's. We now have to show that this instance of the set covering problem, is an appropriate instance of the Dominating Set problem:
Focusing on $F_1, \ldots, F_{K_o}$, we know from Lemma 4 and the fact that the F_i's are formed by polling the G_p's that

$$\exists d_i \in F_i \ : \ d_i \neq v_i \text{and } d_i \in G_{i-1} \quad \text{for } i \ = \ 1, 2, \ldots, K_o.$$

For $i = K_o + 1, \ldots d^*$, pick d_i to be any element of F_i. Let $V_2 = \{d_1, \ldots, d_{d^*}\}$.
We are now ready to define our graph, $G(V, A)$. Let $V = \{1, \ldots N\}$, so that the node labeled i is d_i for $i = 1, 2, ..., d^*$, and is v_i for $i = d^* + 1, \ldots d^* + K_o$. The other nodes are labeled $d^* + K_o + 1, \ldots, N$ in any manner that completes the labeling. Define $g(i) = v_p : i \in G_p$. The neighborhood of these nodes complete the definition:

$$N(i) = \begin{cases} F_i - \{i\}\,, & \text{if } i \ = \ 1, \ldots, d^*; \\ G_{i-1} - \{i\}\,, & \text{if } i \ = \ d^* + 1, \ldots d^* + K_o; \\ g(i)\,, & \text{otherwise;} \end{cases}$$

Observe that this is a valid set of neighborhoods. If we run $GREEDY_{dom}$ on this graph it will pick d_1 through d_{K_o} in the first K_o iterations and all nodes in the optimal dominating set, V_1, will be covered. We ensure this by putting v_i in F_i for $i \leq K_o$ (the boolean *Picked-From-Opt* in the construction procedure does this). At any subsequent iteration, j, v_j will be the least numbered node to cover the maximum number of uncovered nodes (i.e. m_j), and will be picked by the heuristic. Thus, we get the dominating set V_2, of cardinality d^*, for a graph with domination number K_o.
Done.

References

[1] V. Chvatal, "A Greedy Heuristic for the Set Covering problem," *Mathematics of Operations Research, (1979)* vol. 4, pp. 233-235.

[2] D. S. Johnson, "Approximate Algorithms for Combinatorial Problems," *Journal of Computer System Science, (1974)* vol. 9, pp. 256-278.

[3] D. S. Hochbaum, "Approximation Algorithms for the Set Covering and Vertex Covering Problems," *SIAM Journal on Computing, (1982)* vol. 11, pp. 555-556.

[4] L. Lovasz, "On the Ratio of Optimal Integral and Fractional Covers," *Discrete Mathematics,* vol. 13, pp. 383-390.

This research was partially supported by GTE Labs. Waltham, MA.