

The set covering problem (supplemental note)

Kristoffer Arnsfelt Hansen

May 19, 2009

The set covering problem is the following optimization problem.

SET COVER

Instance: Universe X , Family \mathcal{F} of subsets of X .

Objective: Find $\mathcal{C} \subseteq \mathcal{F}$ that minimizes $|\mathcal{C}|$ and satisfies $\cup_{S \in \mathcal{C}} S = X$.

This problem is a generalization of the node covering problem.

NODE COVER

Instance: Graph $G = (V, E)$

Objective: Find $\mathcal{C} \subseteq V$ that minimizes $|\mathcal{C}|$ and satisfies that for all $(u, v) \in E$ either $u \in \mathcal{C}$ or $v \in \mathcal{C}$.

To see this, given an instance $G = (V, E)$ of the NODE COVER problem, we can define a corresponding SET COVER instance as follows.

- $X := E$.
- $\mathcal{F} := \{S_w \mid w \in V\}$, where $S_w = \{(u, v) \in E \mid u = w\}$.

Greedy approximation algorithm

We will study the following approximation algorithm for the set covering problem, that build a set cover by greedily choosing the next set to add that will cover the most new elements.

Input: Universe X , Family \mathcal{F} of subsets of X .

Output: Set cover \mathcal{C} .

```
1:  $U \leftarrow X$ 
2:  $\mathcal{C} \leftarrow \emptyset$ 
3: while  $U \neq \emptyset$  do
4:   pick  $S \in \mathcal{F}$  that maximizes  $|S \cap U|$ 
5:    $U \leftarrow U \setminus S$ 
6:    $\mathcal{C} \leftarrow \mathcal{C} \cup \{S\}$ 
7: end while
8: return  $\mathcal{C}$ 
```

Algorithm 1: Approximation algorithm for SET COVER.

Definition 1 The k th Harmonic number H_k is defined as

$$H_k = 1 + \frac{1}{2} + \cdots + \frac{1}{k} .$$

We have the following simple bound on the Harmonic numbers.

Lemma 2

$$H_k \leq \ln(k) + 1 \leq H_k + 1 .$$

Proof Compare with the integral $\int_1^\infty \frac{1}{x} = \ln(x)$, see Figure 1. □

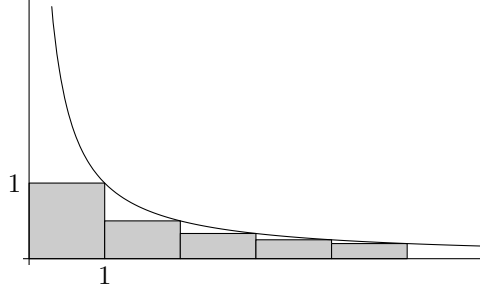


Figure 1: Graph of the function $f(x) = \frac{1}{x}$.

Theorem 3 Algorithm 1 is an polynomial time approximation algorithm for SET COVER with approximation ratio H_k , where $k = \max_{S \in \mathcal{F}} |S|$.

Proof Clearly the algorithm is a polynomial time algorithm; We next give a proof of the approximation ratio.

Let \mathcal{C} be the set cover returned by the algorithm. Let S_1, S_2, \dots, S_m be the sequence of sets added to \mathcal{C} by the algorithm. We will distribute the cost of adding a new set S_i to the cover \mathcal{C} evenly over all new-covered elements.

Assume that element $x \in U$ is covered for the first time in iteration i of the algorithm. We then define the cost of adding x ,

$$c_x = \frac{1}{|S_i \setminus (S_1 \cup \dots \cup S_{i-1})|} .$$

Since at every iteration of the algorithm 1 unit of cost is distributed we have

$$|\mathcal{C}| = \sum_{x \in X} c_x .$$

Now, let \mathcal{C}^* be an optimal set cover. Since \mathcal{C}^* is a set cover, i.e $\cup_{S \in \mathcal{C}^*} S = X$, we have

$$\sum_{x \in X} c_x \leq \sum_{S \in \mathcal{C}^*} \sum_{x \in S} c_x ,$$

and be combining these we have

$$|\mathcal{C}| \leq \sum_{S \in \mathcal{C}^*} \sum_{x \in S} c_x .$$

Let $T \in \mathcal{F}$ be any set of the family \mathcal{F} . Write the elements

$$T = \{y_1, y_2, \dots, y_k\}$$

such that (y_1, y_2, \dots, y_k) is the *reverse* order of when the elements y_1, \dots, y_k are covered by the algorithm.

The central observation is the following: at the moment y_j is in fact covered by the algorithm by set S_i , the set T has at least j elements of X that are not yet covered! Since the algorithm picks the set S_i cover the most new elements, picking S_i must cover the at least j elements (as otherwise the algorithm would have picked T over S_i). In other words we have

$$|S_i \setminus (S_1 \cup \dots \cup S_{i-1})| \geq j ,$$

and it follows

$$c_{y_j} = \frac{1}{|S_i \setminus (S_1 \cup \dots \cup S_{i-1})|} \leq \frac{1}{j} .$$

Hence

$$\sum_{x \in T} c_x \leq \sum_{j=1}^k \frac{1}{j} = H_k .$$

We can now conclude

$$|C| \leq \sum_{S \in \mathcal{C}^*} \left(\sum_{x \in S} c_x \right) \leq \sum_{S \in \mathcal{C}^*} H_{|S|} \leq |\mathcal{C}^*| \cdot \max_{S \in \mathcal{C}^*} H_{|S|} \leq |\mathcal{C}^*| \cdot \max_{S \in \mathcal{F}} H_{|S|} ,$$

and therefore

$$\frac{|C|}{|\mathcal{C}^*|} \leq \max_{S \in \mathcal{F}} H_{|S|}$$

□

Thus we have obtained a $O(\log n)$ approximation algorithm for the set covering problem, by Lemma 2.

Observation 4 *Specializing to VERTEX COVER gives*

$$\frac{|C|}{|\mathcal{C}^*|} \leq \max_{v \in V} H_{\deg(v)} .$$

When the degree of G is at most 3 we have an approximation algorithm with approximation ratio $H_3 = \frac{11}{6} < 2$.