Set Cover

$I = (X, F)$

X = set of objects/elements usually called "Universe"

$F \subseteq \{S : S \subseteq X\} = \mathcal{B}(X)$

"Boolean" of X:
set of all subsets of X

Constraint:

$\forall x \in X \ \exists S \in F : x \in S$

i.e. "F covers X"

optimization problem: find $F' \subseteq F$ s.t.

1) F' covers X
2) $\min |F'|$

Example: $X = \{1, 2, 3, 4, 5\}$

$F = \{\{1,2,3\}, \{2,4\}, \{3,4\}, \{4,5\}\}$

$\Rightarrow F' = \{\{1,2,3\}, \{4,5\}\}$

Applications: - hiring (X = skills, F = people with some skills)

- spamming (X = people, F = mailing lists)

- ....

Set Cover (in its decision version $<(X,F),k>$) is NP-hard.

Proof: vertex cover $<_p$ set cover

$$<G=(V,E),k> \overset{f}{\rightsquigarrow} <(X,F),k>$$

where:

$$X = E$$

$$F = \{S_1, S_2, \dots, S_n\} \quad \text{one } \forall \text{ vertex } 1,2,\dots,n$$

$$S_i = \{e=(u,v) \text{ s.t. } u=i \text{ or } v=i\}$$

now show that finding a set cover of size k $\Longleftrightarrow$ finding a vertex cover of size k

Approx. algorithm: greedy approach

- choose the subset that contains the largest number of uncovered elements

- remove from X those covered elements
- repeat

APPROX_SET_COVER (X, F)

$U \leftarrow X$

$F' \leftarrow \emptyset$ // solution

while $U \neq \emptyset$ do

$\quad$ let $S \in F : |S \cap U| = \max_{S' \in F} \{|S' \cap U|\}$

$\quad U \leftarrow U \setminus S$

$\quad F \leftarrow F \setminus \{S\}$

$\quad F' \leftarrow F' \cup \{S\}$

return F'

Correctness: at every iteration $|U|$ decreases by at least one

Complexity: # of iterations $\leq |X|$

# of iterations $\leq |F|$

$\Rightarrow$ # of iterations $\leq \min\{|X|, |F|\}$

$\forall$ iteration the complexity is $\leq |X||F|$

$$\Rightarrow O\left(|X|\,|F|\,\min\{|X|, |F|\}\right)$$

can be at most cubic in the input size

(with the right data structure can be implemented in $O\left(|X| + |F|\right)$, i.e. linear time)

We'll show that $\dfrac{|F'|}{|F^*|} \leq \lceil \log_2 n \rceil + 1$

OPT

where $n = |X|$

Idea: try to bound the number of iterations such that the set of remaining elements gets empty

$U_0 = X$

U_i = residual universe at the end of the i-th iteration

$|F^*| = k \rightsquigarrow$ unknown

Lemma: after the first k iterations the residual universe at least halved, that is

$$|U_k| \leq \frac{n}{2}$$

$\Rightarrow$ after $k \cdot i$ iterations $|U_{k \cdot i}| \leq \frac{n}{2^i}$

$\Rightarrow$ # of necessary iterations is $\lceil \log_2 n \rceil k + 1$

at each iteration $|F'|$++

$\Rightarrow |F'| \leq \lceil \log_2 n \rceil |F^*| + 1$

$U_k \subseteq X \Rightarrow U_k$ admits a cover of size $\leq k$, all in F (i.e, not selected by the algorithm)

property: if (X, F) admits a cover with $|F| \leq k$ then $\forall X' \subseteq X$ (X', F) admits a cover with $|F| \leq k$

$T_1, T_2, \dots, T_k \in F$ where $\bigcup T_i$ covers U_k

pigeonhole: $\exists \overline{T}$ s.t. $|U_k \cap \overline{T}| \geq \frac{|U_k|}{k}$

we'll now see that in the first k iterations, $\forall$ iteration $|U_k|/k$ elements get covered:

$1 \leq i \leq k \quad S_i \in F$

$\hookrightarrow$ selected subset

$$|S_i \cap U_i| \geq |T_j \cap U_i| \quad \forall 1 \leq j \leq k$$

because T_j has <u>not</u> been selected

this is true for $\overline{T}$, that is

$$|S_i \cap U_i| \geq |\overline{T} \cap U_i| \geq |\overline{T} \cap U_k| \geq \frac{|U_k|}{k}$$

$\Rightarrow$ after the first k iterations the algorithm has covered $\frac{|U_k|}{k} \, k = |U_k|$ elements

$$\underset{\text{covered}}{|U_k|} \leq \underbrace{n - |U_k|}_{\text{residual}}$$

satisfied for $|U_k| \leq \frac{n}{2}$

Is this analysis tight?

Exercise: show that there is an input $I = (X, F)$ on which APPROX_SET_COVER achieves an approximation ratio of $\Theta(\log n)$