# Set Cover

$I = (X, F)$

$X$ = set of elements, called "universe"

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

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

contraint: $\forall x \in X \ \exists S \in F$ s.t. $x \in S$

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 $\langle (X,F), k \rangle$) is NP-hard.

Proof: vertex cover $\leq_p$ set cover

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

where:

$$X = E$$

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

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

now show that finding a set cover of size $k$ $\iff$ 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 until $X = \emptyset$

Approx-Set-Cover $(X, F)$

$U = X$

$F' = \emptyset$ // solution

while $U \neq \emptyset$ do

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

$U = U \setminus S$

$F = F \setminus \{S\}$

$F' = F' \cup \{S\}$

return $F'$

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

Complexity : $n^\circ$ of iterations $\leq |X|$

$n^\circ$ of iterations $\leq |F|$

$\Rightarrow$ $n^\circ$ of iterations $\leq \min\{|X|, |F|\}$

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

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

is at most cubic in the input size

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

We'll now 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 <u>remaining</u> <u>elements</u> gets empty

$U_0 = X$

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

$|F^*| = k$ unknown

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

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

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

$\implies$ n° of necessary iterations is $\lceil \log_2 n \rceil k + 1$

at each iteration $|F'|$++

$$\implies |F'| \leq \lceil \log_2 n \rceil \underset{\shortparallel}{k} + 1$$

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

Let's prove the Lemma:

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

(trivial) 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 $\cup T_i$ covers $U_k$

pigeonhole principle: $\exists \bar{T}$ s.t. $|U_k \cap \bar{T}| \geq \frac{|U_k|}{K}$

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

$1 \leq i \leq K$ $S_i \in F$ selected subsets

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

because $T_j$ has not been selected $\quad |U_i| \geq |U_k|$

this holds for $\bar{T}$, that is

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

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

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

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

Is this analysis tight?

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

(Hint: the alg. chooses the set that contains the largest n° of uncovered elements, whereas OPT chooses a set that contains the second-largest n° of uncovered elements)