Approximation Algorithms 3: Set Cover and Hitting Set

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong

Set Cover

Let U be a finite set called the **universe**.

We are given a family $\mathcal{S}$ where

- each member of $\mathcal{S}$ is a set $S \subseteq U$;
- $\bigcup_{S \in \mathcal{S}} S = U$.

A sub-family $\mathcal{C} \subseteq \mathcal{S}$ is a **universe cover** if every element of U appears in at least one set in $\mathcal{C}$.

- Define the **cost** of $\mathcal{C}$ as $|\mathcal{C}|$.

The set cover problem:
Find a universe cover with the smallest cost.

Example: $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and $\mathcal{S} = \{S_1, S_2, ..., S_5\}$ where

$$\begin{aligned} S_1 &= \{1, 2, 3, 4\} \\ S_2 &= \{2, 5, 7\} \\ S_3 &= \{6, 7\} \\ S_4 &= \{1, 8\} \\ S_5 &= \{1, 2, 3, 8\}. \end{aligned}$$

An optimal solution is $\mathcal{C} = \{S_1, S_2, S_3, S_4\}$.

The input size of the set cover problem is $n = \sum_{S \in \mathcal{S}} |S|$.

The problem is NP-hard.

- No one has found an algorithm solving the problem in time polynomial in n.
- Such algorithms cannot exist if $\mathcal{P} \neq \mathcal{NP}$.

$\mathcal{A}$ = an algorithm that, given any legal input $\mathcal{S}$ with universe U, returns a universe cover $\mathcal{C}$.

Denote by $OPT_{\mathcal{S}}$ the smallest cost of all universe covers when the input family is $\mathcal{S}$.

> $\mathcal{A}$ is a ρ-**approximate algorithm** for the set cover problem if, for any legal input $\mathcal{S}$, $\mathcal{A}$ can return a universe cover with cost at most $\rho \cdot OPT_{\mathcal{S}}$.

The value ρ is the **approximation ratio**.
We say that $\mathcal{A}$ achieves an approximation ratio of ρ.

Consider the following algorithm.

Input: A family $\mathcal{S}$

1. $\mathcal{C} = \emptyset$
2. **while** U still has elements not covered by any set in $\mathcal{C}$
3. $\quad F \leftarrow$ the set of elements in U not covered by any set in $\mathcal{C}$
   /* for each set $S \in \mathcal{S}$, define its **benefit** to be $|S \cap F|$ */
4. $\quad$ add to $\mathcal{C}$ a set in $\mathcal{S}$ with the largest benefit
5. **return** $\mathcal{C}$

It is easy to show:

- The $\mathcal{C}$ returned is a universe cover;
- The algorithm runs in time polynomial to n.

We will prove later that the algorithm is $(1 + \ln |U|)$-approximate.

Example: $S_1 = \{1, 2, 3, 4\}$, $S_2 = \{2, 5, 7\}$, $S_3 = \{6, 7\}$, $S_4 = \{1, 8\}$, $S_5 = \{1, 2, 3, 8\}$

- In the beginning, $\mathcal{C} = \emptyset$ and $F = \{1, 2, 3, 4, 5, 6, 7, 8\}$.
- Next, we can add S_1 or S_5 to $\mathcal{C}$ (benefit 4). The choice is arbitrary; suppose we add S_1. Now, $F = \{5, 6, 7, 8\}$.
- Next, we can add S_2 or S_3 (benefit 2). The choice is arbitrary; suppose we add S_2. Now, $F = \{6, 8\}$.
- Next, we can add S_3, S_4, or S_5 (benefit 1). The choice is arbitrary; suppose we add S_3. Now, $F = \{8\}$.
- Next, we cab add S_4 or S_5 (benefit 1). The choice is arbitrary; suppose we add S_4. Now, $F = \emptyset$.

The algorithm terminates with $\mathcal{C} = \{S_1, S_2, S_3, S_4\}$.

Theorem 1: The algorithm returns a universe cover with cost at most $1 + (\ln |U|) \cdot OPT_{\mathcal{S}} \leq (1 + \ln |U|) \cdot OPT_{\mathcal{S}}$.

$\mathcal{C}$ = the universe cover returned.
$t = |\mathcal{C}|$.

Denote the sets in $\mathcal{C}$ as $S_1, S_2, ..., S_t$, picked in the order shown.

For each $i \in [1, t]$, define z_i as the size of F after S_i is picked.
Specially, define $z_0 = |U|$.

> $z_t = 0$ and $z_{t-1} \geq 1$. **Think:** why?

Denote by $\mathcal{C}^*$ an optimal universe cover, namely, $OPT_{\mathcal{S}} = |\mathcal{C}^*|$.

We will prove later:

Lemma 1: For $i \in [1, t]$, it holds that

$$z_i \leq z_{i-1} \cdot \left(1 - \frac{1}{OPT_{\mathcal{S}}}\right).$$

From Lemma 1, we get:

$$
\begin{aligned}
z_{t-1} &\le z_{t-2} \cdot \left(1 - \frac{1}{OPT_{\mathcal{S}}}\right) \\
&\le z_{t-3} \cdot \left(1 - \frac{1}{OPT_{\mathcal{S}}}\right)^2 \\
&\ldots \\
&\le z_0 \cdot \left(1 - \frac{1}{OPT_{\mathcal{S}}}\right)^{t-1} = |U| \cdot \left(1 - \frac{1}{OPT_{\mathcal{S}}}\right)^{t-1} \\
&\le |U| \cdot e^{-\frac{t-1}{OPT_{\mathcal{S}}}}
\end{aligned}
$$

where the last inequality used the fact $1 + x \le e^x$ for any real value x.

As $z_{t-1} \ge 1$, we have

$$1 \le |U| \cdot e^{-\frac{t-1}{OPT_{\mathcal{S}}}} \quad (1)$$

which resolves to $t \le 1 + (\ln |U|) \cdot OPT_{\mathcal{S}}$. This proves Theorem 1.

Proof of Lemma 1

Before z_i is chosen, F has z_{i-1} elements.

At this moment, at least one set in $\mathcal{C}^*$ has a benefit at least $\frac{z_{i-1}}{|\mathcal{C}^*|} = \frac{z_{i-1}}{OPT_{\mathcal{S}}}$ (every element of F must appear in some set in $\mathcal{C}^*$).

Hence, S_i must have a benefit at least $\frac{z_{i-1}}{OPT_{\mathcal{S}}}$ (greedy). Therefore:

$$\begin{aligned} z_i &= |F \setminus S_i| = |F| - |F \cap S_i| \\ &\le z_{i-1} - \frac{z_{i-1}}{OPT_{\mathcal{S}}} \\ &= z_{i-1}\left(1 - \frac{1}{OPT_{\mathcal{S}}}\right) \end{aligned}$$

□

Next, we will introduce a closely related problem called the **hitting set problem**.

Hitting Set

Let U be a finite set called the **universe**.

We are given a family $\mathcal{S}$ where

- each member of $\mathcal{S}$ is a set $S \subseteq U$;
- $\bigcup_{S \in \mathcal{S}} S = U$.

A subset $H \subseteq U$ **hits** a set $S \in \mathcal{S}$ if $H \cap S \neq \emptyset$.
A subset $H \subseteq U$ is a **hitting set** if it hits all the sets in $\mathcal{S}$.

The hitting set problem:
Find a hitting set H of the minimize size.

Example: $U = \{1, 2, 3, 4, 5\}$ and $\mathcal{S} = \{S_1, S_2, ..., S_8\}$ where

$$\begin{aligned} S_1 &= \{1, 4, 5\} \\ S_2 &= \{1, 2, 5\} \\ S_3 &= \{1, 5\} \\ S_4 &= \{1\} \\ S_5 &= \{2\} \\ S_6 &= \{3\} \\ S_7 &= \{2, 3\} \\ S_8 &= \{4, 5\} \end{aligned}$$

An optimal solution is $H = \{1, 2, 3, 4\}$.

The input size of the set cover problem is $n = \sum_{S \in \mathcal{S}} |S|$.

The problem is NP-hard.

- No one has found an algorithm solving the problem in time polynomial in n.
- Such algorithms cannot exist if $\mathcal{P} \neq \mathcal{NP}$.

$\mathcal{A}$ = an algorithm that, given any legal input $\mathcal{S}$ with universe U, returns a hitting set.

Denote by $OPT_{\mathcal{S}}$ the smallest size of all hitting sets.

$\mathcal{A}$ is a ρ-**approximate algorithm** for the hitting set problem if, for any legal input $\mathcal{S}$, $\mathcal{A}$ can return a hitting set with size at most $\rho \cdot OPT_{\mathcal{S}}$.

The value ρ is the **approximation ratio**.
We say that $\mathcal{A}$ achieves an approximation ratio of ρ.

We can convert the hitting set problem to set cover.

Let $(U_{hs}, \mathcal{S}_{hs})$ be the input to the hitting set problem. W.l.o.g., assume that $\mathcal{S}_{hs} = \{S_1, S_2, ..., S_t\}$.

We create an instance of the set cover problem as follows:

- $U_{sc} = \{1, 2, ..., t\}$.
- For each element $e \in U_{hs}$, define $OriginS_e = \{i \mid 1 \le i \le t \text{ and } e \in S_i\}$.
- Then, create $\mathcal{S}_{sc} = \{OriginS_e \mid e \in U_{hs}\}$.

Theorem 2: $(U_{hs}, \mathcal{S}_{hs})$ has a hitting set of size s if and only if $(U_{sc}, \mathcal{S}_{sc})$ has a universe cover of size s.

We therefore have a polynomial-time algorithm solving the hitting set problem with approximation ratio $1 + \ln U_{sc} = 1 + \ln t \leq 1 + \ln n$.

Next we will prove the theorem.

Proof of the $\Rightarrow$ Direction: Namely, if $(U_{hs}, \mathcal{S}_{hs})$ has a hitting set of size s, then $(U_{sc}, \mathcal{S}_{sc})$ has a universe cover of size s.

Let H be any hitting set. Construct

$$\mathcal{C}_H = \{OriginS_e \mid e \in H\}.$$

We argue that $\mathcal{C}_H$ is a universe cover for $(U_{sc}, \mathcal{S}_{sc})$.

Suppose that this is not true. Then, there is an integer $i \in [1, t]$ that does not belong to $\mathcal{C}_H$. This means that $i \notin OriginS_e$ for any $e \in H$. Hence, S_i does not contain any element in H. This contradicts H being a hitting set.

Proof of the $\Leftarrow$ Direction: Namely, if $(U_{sc}, \mathcal{S}_{sc})$ has a universe cover of size s, then $(U_{hs}, \mathcal{S}_{hs})$ has a hitting set of size s.

Let $\mathcal{C}$ be any universe cover. Construct

$$H_{\mathcal{C}} = \{e \mid OriginS_e \in \mathcal{C}\}.$$

We argue that $H_{\mathcal{C}}$ is a hitting set for $(U_{hs}, \mathcal{S}_{hs})$.

Suppose that this is not true. Then, $\mathcal{S}_{hs}$ has an S_i — for some integer $i \in [1, t]$ — that contains no elements in $H_{\mathcal{C}}$. This means that $i \notin OriginS_e$ for any $e \in H_{\mathcal{C}}$. Because $\mathcal{C} = \{OriginS_e \mid e \in H_{\mathcal{C}}\}$, we conclude that i does not appear in any set of $\mathcal{C}$. This contradicts $\mathcal{C}$ being a universe cover. □