Approximation Algorithms II

2023/11/23

詹博华（中国科学院软件研究所）

Set-covering problem

- Given a finite set X and a family F of subsets of X, such that every element of X belongs to at least one subset in F:

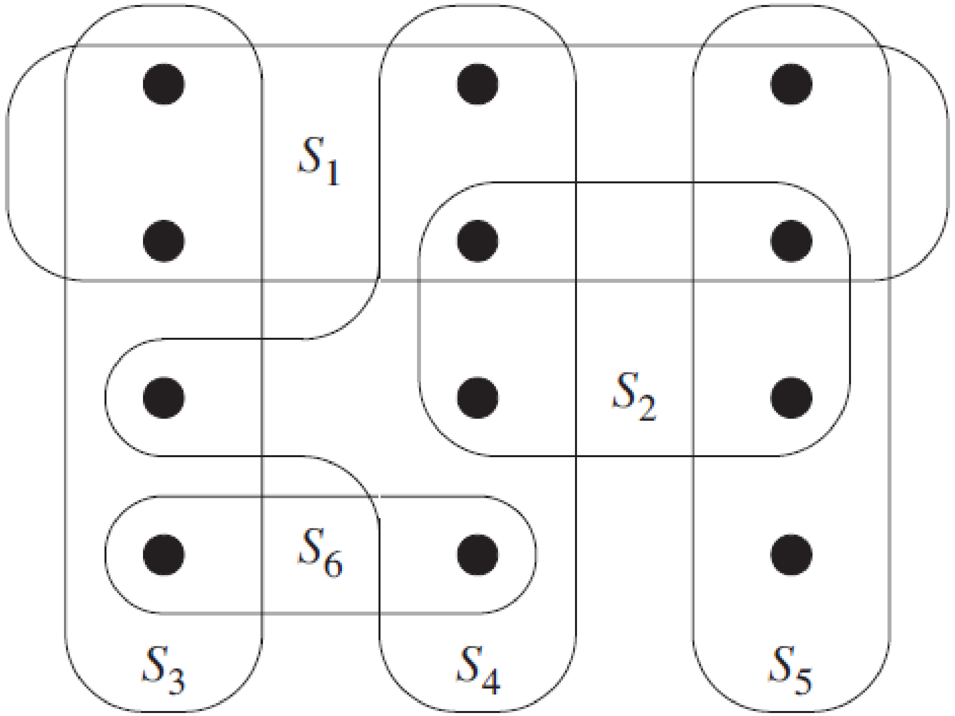
$$X = \bigcup_{S \in F} S$$

- Find a subset $C \subseteq F$ of minimum size such that C still covers X. That is:

$$X = \bigcup_{S \in C} S$$

Set-covering problem: example

- The set X consists of 12 points.
- The family F consists of six subsets of X: $\{S_1, S_2, S_3, S_4, S_5, S_6\}$.
- A cover of minimum size is $\{S_3, S_4, S_5\}$, with three subsets.



Set-covering is NP-complete

- An easy reduction from vertex-cover problem.
- Given graph $G = (V, E)$, let S be the set of edges E. For each vertex v, construct a subset S_v as the set of edges incident on v. Then let $F = \{S_v | v \in V\}$. Any subset of F covering S corresponds to a vertex-cover of G with the same size.

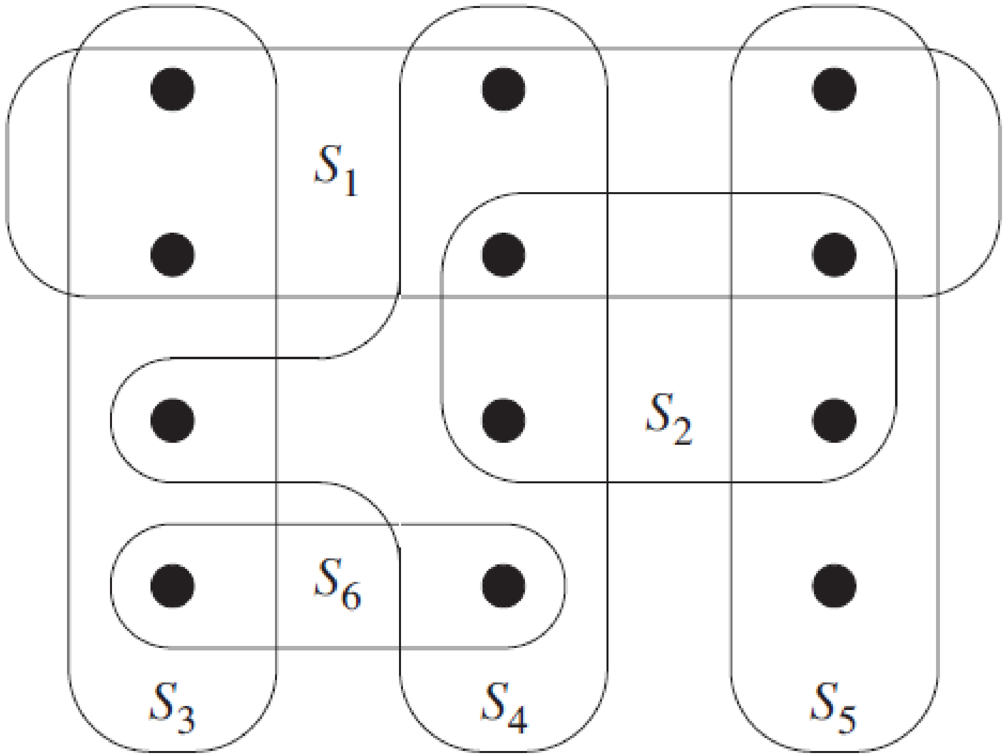
A greedy approximation algorithm

- At each stage, pick the set S that covers the greatest number of remaining elements that are uncovered.

```
GREEDY-SET-COVER(X, ℱ)
U = X
𝒞 = ∅
while U ≠ ∅
    select an S ∈ ℱ that maximizes |S ∩ U|
    U = U − S
    𝒞 = 𝒞 ∪ {S}
return 𝒞
```

Greedy approximation algorithm: example

- With the problem given on the right, the greedy algorithm will first choose S_1 (covering 6 points), then S_4 (covering another 3 points), then S_5 (covering another 2 points). Finally choose either S_3 or S_6 to cover the remaining point.
- This gives set-covering with four subsets.



Analysis

- Define the harmonic number $H(d)$ as:

$$H(d) = \sum_{i=1}^{d} 1/i$$

- We have $H(d)$ increases logarithmically with d.
- **Theorem:** the greedy algorithm has approximation ratio ρ, where

$$\rho = H(\max\{|S|: S \in F\}),$$

  that is, harmonic number of the size of the largest subset in F.
- The proof shown as follows is quite technical.

Some definitions

- Let C be the cover returned by the greedy algorithm.
- Let C^* be the optimal set-covering.
- Let S_i be the i^{th} subset selected by the greedy algorithm.
- We spread the *cost* of S_i among the elements first covered by S_i. That is, let c_x denote the cost allocated to element x, defined by

$$c_x = \frac{1}{|S_i - (S_1 \cup S_2 \cup \cdots \cup S_{i-1})|}$$

- Then the total cost is

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

Illustration of definitions so far

- Rephrase the example as follows: given 12 points $x_1, \ldots, x_{12}$, and the following subsets:

$$\{x_1, x_2, x_3, x_4, x_5, x_6\},\ \{x_5, x_6, x_8, x_9\},\ \{x_1, x_4, x_7, x_{10}\},$$
$$\{x_2, x_5, x_7, x_8, x_{11}\},\ \{x_3, x_6, x_9, x_{12}\},\ \{x_{10}, x_{11}\}.$$

- The subsets picked are (bold indicate new points):
  - $S_1 = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \boldsymbol{x_4}, \boldsymbol{x_5}, \boldsymbol{x_6}\}$
  - $S_2 = \{x_2, x_5, \boldsymbol{x_7}, \boldsymbol{x_8}, \boldsymbol{x_{11}}\}$
  - $S_3 = \{x_3, x_6, \boldsymbol{x_9}, \boldsymbol{x_{12}}\}$
  - $S_4 = \{\boldsymbol{x_{10}}, x_{11}\}$

The assigned cost c_x are:
$c_1 = c_2 = c_3 = c_4 = c_5 = c_6 = 1/6$,
$c_7 = c_8 = c_{11} = 1/3$,
$c_9 = c_{12} = 1/2$,
$c_{10} = 1$.

Analysis continued

- **Crucial observation:** consider the optimal set covering C^*. Since the subsets in C^* covers each element in X at least once, we have:

$$\sum_{S \in C^*} \sum_{x \in S} c_x \geq \sum_{x \in X} c_x = |C|$$

- Hence, it is of interest to give an upper bound on the sum $\sum_{x \in S} c_x$ for any subset $S \in F$.
- **Main Lemma:**

$$\sum_{x \in S} c_x \leq H(|S|)$$

for any S belonging to F.

Analysis continued

- First, we finish the proof assuming the main lemma.
- Then
$$|C| \leq \sum_{S \in C^*} \sum_{x \in S} c_x \leq \sum_{S \in C^*} H(|S|) \leq |C^*| \cdot H(\max\{|S|: S \in F\})$$
- This prove the approximation ratio of $H(\max\{|S|: S \in F\})$.

Proof of Main Lemma

We now continue with proof of the main lemma.

- Given $S \in F$, consider how it is covered by each of the subsets S_i picked by the greedy algorithm. Let
$$u_i = |S - (S_1 \cup S_2 \cup \cdots \cup S_i)|.$$
- That is, u_i is the number of elements left uncovered after the i^{th} iteration of the greedy algorithm.
- We have $u_0 = |S|$, and $u_{i-1} - u_i$ is the number of elements newly covered by S_i. So we have:
$$\sum_{x \in S} c_x = \sum\nolimits_{i=1}^{k} (u_{i-1} - u_i) \cdot \frac{1}{|S_i - (S_1 \cup S_2 \cup \cdots \cup S_{i-1})|}$$

Proof of Main Lemma: Example

Consider the set $\mathrm{S} = \{x_5, x_6, x_8, x_9\}$ which is not picked. We have:

- x_5, x_6 is covered by $S_1 = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \boldsymbol{x_4}, \boldsymbol{x_5}, \boldsymbol{x_6}\}$.
- x_8 is covered by $S_2 = \{x_2, x_5, \boldsymbol{x_7}, \boldsymbol{x_8}, \boldsymbol{x_{11}}\}$.
- x_9 is covered by $S_3 = \{x_3, x_6, \boldsymbol{x_9}, \boldsymbol{x_{12}}\}$.
- So $u_0 = 4$, $u_1 = 2$, $u_2 = 1$, $u_3 = 0$, and

$$\sum_{x \in S} c_x = (u_0 - u_1) \cdot \frac{1}{6} + (u_1 - u_2) \cdot \frac{1}{3} + (u_2 - u_3) \cdot \frac{1}{2} = 2 \cdot \frac{1}{6} + \frac{1}{3} + \frac{1}{2}$$

Proof of Main Lemma, Step 2

- **Crucial observation #2:** for each i, we have the inequality:
  $|S_i - (S_1 \cup S_2 \cup \cdots \cup S_{i-1})| \geq |S - (S_1 \cup S_2 \cup \cdots \cup S_{i-1})| = u_{i-1}$.
- This says: there is at least as many elements in S_i uncovered by $S_1, \dots, S_{i-1}$ as there are elements in S uncovered by $S_1, \dots, S_{i-1}$. This has to hold, for otherwise S will be picked in the greedy algorithm rather than S_i.
- So we have the inequality:

$$\sum_{x \in S} c_x \leq \sum_{i=1}^{k} (u_{i-1} - u_i) \cdot \frac{1}{u_{i-1}}$$

Proof of Main Lemma: Example

- We continue the example two slides before:

$$\begin{aligned}\sum_{x\in S} c_x &= (u_0 - u_1)\cdot\frac{1}{6} + (u_1 - u_2)\cdot\frac{1}{3} + (u_2 - u_3)\cdot\frac{1}{2} \\ &\le (u_0 - u_1)\cdot\frac{1}{u_0} + (u_1 - u_2)\cdot\frac{1}{u_1} + (u_2 - u_3)\cdot\frac{1}{u_2} \\ &= (4-2)\cdot\frac{1}{4} + (2-1)\cdot\frac{1}{2} + (1-0)\cdot\frac{1}{1}\end{aligned}$$

Proof of Main Lemma, Step 3

- **Crucial observation #3:** the sum

$$\sum_{i=1}^{k}(u_{i-1}-u_i)\cdot\frac{1}{u_{i-1}}$$

  is bounded by the harmonic series.

- To give an example:

$$(4-2)\cdot\frac{1}{4}+(2-1)\cdot\frac{1}{2}+(1-0)\cdot\frac{1}{1}=\frac{1}{4}+\frac{1}{4}+\frac{1}{2}+\frac{1}{1}$$
$$\leq\frac{1}{4}+\frac{1}{3}+\frac{1}{2}+\frac{1}{1}$$

Proof of Main Lemma: Step 3

- In general:

$$\sum_{x \in S} c_x \leq \sum_{i=1}^{k} (u_{i-1} - u_i) \cdot \frac{1}{u_{i-1}} \leq \sum_{i=1}^{|S|} \frac{1}{i} = H(|S|)$$

- This finishes proof of the Main Lemma.
- **Conclusion:** greedy algorithm approximates set-covering within a factor of $H(d)$, where d is the size of the largest subset in F.
- This can give quite good theoretical results if d is small. For example, for vertex-cover on a graph where degree of each vertex is at most 3, this gives approximation ratio of $H(3) = 11/6$, which is better than the ratio 2 given earlier.