27 Oct 2021

Greedy Set Cover

Set cover: universe $\mathcal{U} = \{u_1, \ldots, u_n\}$

subsets $S_1, S_2, \ldots, S_m$ $\bigcup_{i=1}^{m} S_i = \mathcal{U}$.

costs $c_i > 0$ (cost set S_i)

Question: Choose a subcollection $\{S_i \mid i \in J\}$ whose union is $\mathcal{U}$. Minimize total cost $\sum_{i \in J} c_i$.

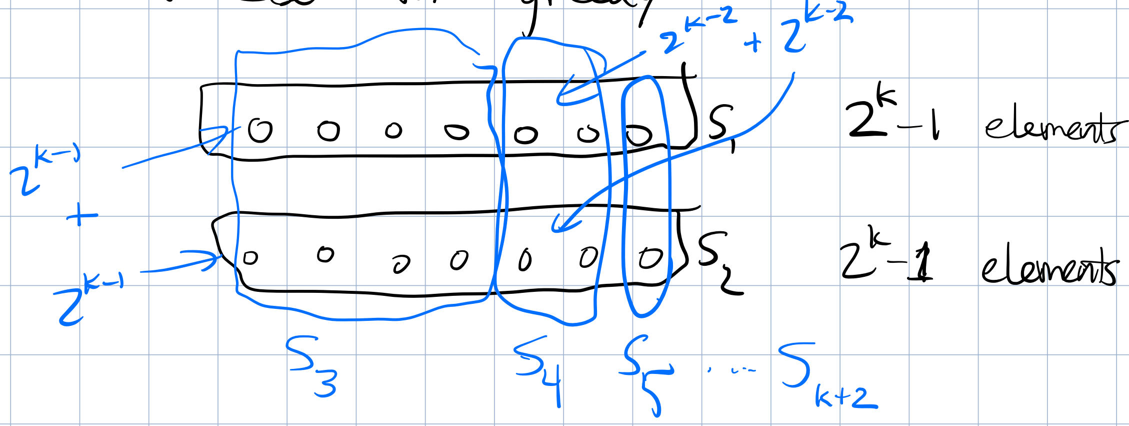
Vertex cover: $\mathcal{U} = \{\text{edges of } G\}$.

$S_i = \{\text{edges incident to } v_i\}$

$c_i = w(v_i)$

Greedy algorithm: Start by choosing S_i where i maximizes $|S_i|/c_i$.
Modify $\mathcal{U}$ by deleting elements of S_i.
While $\mathcal{U}$ is non-empty, recursively call greedy algorithm to cover it.

Bad case for greedy:



S_{2+j} for $j > 1$ has $2^{k-j} + 2^{k-j}$ elements.

$c_i = 1 \quad \forall i.$

Greedy picks k sets when 2 would suffice.

$\Longrightarrow$ approximation is $k/2$

$\approx \frac{1}{2} \log(n).$

Analyzing Greedy approx factor.

Assume for now $c_i = 1 \ \forall i.$

Let k denote the # sets in the optimum solution.

W.l.o.g. we'll assume $\mathcal{U} = S_1 \cup \dots \cup S_k.$

Plan of attack: Let $n(t) :=$ # of uncovered elements after t iterations of greedy alg. Try to show $n(t)$ decreases rapidly as a function of t.

In any iteration, $t+1$, we know all the remaining uncovered elements are covered by $S_1 \cup \dots \cup S_k$, so at least one of $S_1, \dots, S_k$ has $\geq \frac{1}{k} \cdot n(t)$ of the uncovered elements.

$\implies$ Greedy selects a set that covers $\frac{1}{k} n(t)$ or more of the uncovered elements.

$$\implies \quad n(t+1) \leq \left(1 - \frac{1}{k}\right) \cdot n(t).$$

We know $n(0) = n = |\mathcal{U}|$.
So
$$n(t) \leq \left(1 - \frac{1}{k}\right)^t \cdot n.$$
If t is large enough that $\left(1-\frac{1}{k}\right)^t < \frac{1}{n}$, then $n(t) = 0$ so greedy takes at most t iterations.

Now use $1 - x \le e^{-x}$.

So if $t > k \ln(n)$

$$\left(1-\frac{1}{k}\right)^t < e^{-(1/k)\cdot t} < e^{-\ln(n)} = \frac{1}{n}.$$

Conclusion: when $c_i = 1 \ \forall i$

and optimum value of SET COVER is k

then greedy solution uses

$$\le \lceil k \ln(n) \rceil$$

sets.

BONUS. In max. coverage problem we are given $\mathcal{U} = \{u_1, \ldots, u_n\}$ and $S_1, S_2, \ldots, S_m$ with $S_1 \cup \ldots \cup S_m = \mathcal{U}$ and budget parameter k.

Goal. Choose k sets whose union is as large as possible.

can be modified to show

The analysis above ~~shows~~ Greedy computes a $\left(1 - \frac{1}{e}\right)$-approx to the optimum value.

redefine $n(t)$ to be the diff

btw # elements covered after

t iterations and # elements covered by the best k sets.

LP relaxation of set cover:

$$\min \sum_{i=1}^{m} c_i x_i$$

decision variable representing picking set S_i

$$\text{s.t.} \sum_{i:\, u_j \in S_i} x_i \geq 1 \qquad \forall j$$

$$x_i \geq 0 \qquad \forall i$$

Dual of LP:

$$\max \sum_{j=1}^{n} y_j$$

$$\text{s.t.} \sum_{j \in S_i} y_j \leq c_i \qquad \forall i=1,\dots,m$$

$$y_j \geq 0$$

Game plan: when greedy picks S_i containing q uncovered elements $\{u_{j_1}, \dots, u_{j_q}\}$ divide its cost equally:

$$z_j = \frac{1}{q} \cdot c_i \quad \text{for } j \in \{\hat{j}_1, \ldots, \hat{j}_q\}.$$

We'll see this vector $\vec{z}$ doesn't satisfy dual LP, but

$$y = \frac{1}{\alpha} \cdot \vec{z}$$

satisfies the dual for $\alpha \approx \ln(n)$.

For any set S_i number its elements as $\{u_1, u_2, \ldots, u_r\}$ where $r = |S_i|$ and elements are numbered in decreasing order of the iteration when Greedy covered them.

$$z_{u_1} \leq c_i$$

$$z_{u_2} \leq c_i/2$$

$$\vdots$$

$$z_{u_t} \leq c_i/t$$

$$\sum_{j \in S_i} z_j \leq c_i \left(1 + \frac{1}{2} + \cdots + \frac{1}{t}\right)$$

$$< c_i \left(1 + \ln(t)\right)$$

Conclusion: Greedy approx ratio is $\leq 1 + \ln\left(\max_i |S_i|\right)$.

(This method of analysis is called "dual fitting".)