Greedy Set Cover

Set cover: universe $U = \{ u_1, \ldots, u_n \}$
subsets $S_1, S_2, \ldots, S_m$ $\cup_{i=1}^m S_i = U$

$\text{costs } c_i \geq 0$ (cost set $S_i$)

Question: Choose a subcollection $\{ S_i \mid i \in J \}$ whose union is $U$.

Minimize total cost $\sum_{i \in J} c_i$.

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

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

$C_i = w(v_i)$

Greedy algorithm: Start by choosing $S_i$ where $i$ maximizes $|S_i|/c_i$.

Modify $U$ by deleting elements of $S_i$.

While $U$ is non-empty, recursively call greedy algorithm to cover it.
Bad case for greedy:

\[ S_1 \cup S_1 \cup S_2 \cup \ldots \cup S_{k+2} \]

\[ 2^{k-1} \text{ elements} \]

\[ S_{2+j} \text{ for } j > 1 \text{ has } 2^k j + 2^{k-j} \text{ elements.} \]

\[ C_i = 1 \quad \forall i \]

Greedy picks \( k \) sets when \( 2 \) would suffice.

\[ \Rightarrow \text{ approximation is } \frac{k}{2} \]

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

Analyzing Greedy approx factor.

Assume for now \( C_i = 1 \quad \forall i \).

Let \( k \) denote the \# sets in the optimum solution.

Wlog, we'll assume \( U = S_1 \cup \ldots \cup S_k \).
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 \( S_1, \ldots, S_k \), so at least one of \( S_1, \ldots, S_k \) has \( \geq \frac{1}{k} \cdot n(t) \) of the uncovered elements.

\[ \implies \text{Greedy selects a set that covers } \geq \frac{1}{k} \cdot n(t) \text{ of the uncovered elements.} \]

\[ \implies n(t+1) \leq (1 - \frac{1}{k}) \cdot n(t). \]

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

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

\[
(1 - \frac{1}{t}) < e^{-(\frac{1}{t})} t < e^{-\ln(n)} = \frac{1}{n}.
\]

Conclusion: when \( c_i = 1 \ \forall i \) and optimum value of SET COVER is \( k \) then greedy solution uses

\[
\leq \left\lceil k \ln(n) \right\rceil
\]

sets.

Bonus. In max. coverage problem we are given \( U = \{ u_1, \ldots, u_n \} \) and \( S_1, S_2, \ldots, S_m \) with \( S_1 \cup \cdots \cup S_m = U \) and budget parameter \( k \).

Goal: Choose \( k \) sets whose union is as large as possible.

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

redefine \( n(t) \) to be the diff.

\( \# \) elements covered after
to iterations and # elements covered by the best $k$ sets.

**LP relaxation of set cover:**

\[
\begin{align*}
\min & \quad \sum_{i=1}^{m} c_i x_i & \quad & \text{decision variable representing picking set } S_i \\
\text{st.} & \quad \sum_{i : j \in S_i} x_i \geq 1 & \quad & \forall j \\
& \quad x_i \geq 0 & \quad & \forall i
\end{align*}
\]

**Dual of LP:**

\[
\begin{align*}
\max & \quad \sum_{j=1}^{n} y_j \\
\text{st.} & \quad \sum_{j \in S_i} y_j \leq c_i & \quad & \forall i = 1, \ldots, m \\
& \quad y_j \geq 0
\end{align*}
\]

Game plan: when greedy picks $S_i$ containing $u_j, \ldots, y_{j_0}$ uncovered elements divide its cost equally:
\[ \mathbf{z}_j = \frac{1}{\theta} \cdot \mathbf{c}_i \text{ for } j \in \{j_1, \ldots, j_2\}. \]

We'll see this vector \( \mathbf{z} \) doesn't satisfy dual \( L_\frac{1}{2} \), but \( \mathbf{y} = \frac{1}{\lambda^2} \cdot \mathbf{z} \) 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.

\[ \sum_{u_1} \leq C_i \]
\[ \sum_{u_2} \leq C_i/2 \]
\[ \vdots \]
\[ \sum_{u_r} \leq C_i/2 \]
\[ \sum_{j \in S_i} z_j \leq c_i \left( 1 + \frac{1}{2} + \ldots + \frac{1}{t} \right) \]
\[ < c_i \left( 1 + \ln(t) \right) \]

Conclusion, Greedy approx ratio
\[ \frac{z}{z^*} \leq 1 + \ln(\max_i |S_i|) \].

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