Cornell University, Fall 2020	CS 6820: Algorithms
Lecture notes: Submodular functions	2 Dec 2020

If Ω is a set and 2^Ω denotes its power set, a function $f : 2^\Omega \to \mathbb{R}$ is called *submodular* if it satisfies

$$\forall A, B \subseteq \Omega \quad f(A) + f(B) \geq f(A \cup B) + f(A \cap B). \tag{1}$$

More generally, if L is a lattice then $f : L \to \mathbb{R}$ is called *submodular* if

$$\forall a, b \in L \quad f(a) + f(b) \geq f(a \vee b) + f(a \wedge b). \tag{2}$$

In these notes we shall limit ourselves to considering the special case in which $L = 2^\Omega$ for a finite set Ω.

When Ω is finite, the following equivalent definition of submodularity lends insight into its meaning. A function $f : 2^\Omega \to \mathbb{R}$ is submodular if and only if it satisfies the following "diminishing marginal returns" property:

$$\forall S \subseteq T \subset \Omega,\, i \in \Omega \setminus T \quad f(S \cup \{i\}) - f(S) \geq f(T \cup \{i\}) - f(T). \tag{3}$$

This relation is easily seen to follow from property (1) by substituting $A = S \cup \{i\}$, $B = T$. Conversely, property (3) implies property (1) by rewriting the latter property as

$$\forall A, B \subseteq \Omega \;\; f(A) - f(A \cap B) \geq f(A \cup B) - f(B)$$

and deriving that inequality from (3) by induction on the cardinality of $A \setminus B$.

Submodular functions are pervasive in combinatorial optimization, and they arise quite often in economic theory (where they may represent a consumer's value for receiving a bundle of goods) and machine learning (where they sometimes arise as the log-probability function in certain binary labeling problems, and they can also be used to model summarization and feature selection tasks). These notes cover some basic structural and algorithmic facts about submodular functions.

1 Examples

In this section we collect some examples of submodular functions, to illustrate the many contexts in which they arise.

Example 1 (Constant functions)**.** A function that satisfies $f(A) = f(\emptyset)$ for all $A \subseteq \Omega$ is trivially submodular.

Example 2 (Additive functions)**.** The simplest non-trivial example of a submodular function is an additive set function, i.e. one which satisfies

$$f(A \cup B) = f(A) + f(B)$$

for every pair of disjoint sets $A, B \subseteq \Omega$. Every such function is a "weighted cardinality function", i.e. there is a function $w : \Omega \to \mathbb{R}$ — given by $w(i) = f(\{i\})$ — such that $f(A) = \sum_{i \in A} w(i)$.

Example 3 (Modular functions)**.** A function $f : 2^\Omega \to \mathbb{R}$ is *modular* if it satisfies

$$f(A) + f(B) = f(A \cup B) + f(A \cap B)$$

for all $A, B \subseteq \Omega$. Modular functions are obviously submodular. It is not difficult to check that a function is modular if and only if it is the sum of a constant function and an additive function.

Example 4 (Concave functions of cardinality)**.** If $h : \mathbb{N} \to \mathbb{R}$ is a concave function, meaning that $h(k+1) - h(k)$ is a non-increasing function of $k \in \mathbb{N}$, then the function $f(A) = h(|A|)$ is submodular, because it satisfies (3).

Example 5 (Coverage functions)**.** Suppose we are given a collection of finite sets $\{S_i : i \in \Omega\}$. For $A \subseteq \Omega$ let

$$S_A = \bigcup_{i \in A} S_i, \qquad f(A) = |S_A|$$

where $|\cdot|$ denotes the cardinality of a set. In other words, $f(A)$ is the number of elements covered by the sets indexed by A. This function is submodular because for all $A \subseteq B$ and $i \notin B$, $f(A \cup \{i\}) - f(A) = |S_i \setminus S_A|$ while $f(B \cup \{i\}) - f(B) = |S_i \setminus S_B|$. The latter quantity cannot be greater than the former, because $S_B \supseteq S_A$.

Example 6 (Weighted coverage functions)**.** The previous example generalizes to the case in which we have a measure space (X, μ), a collection of measurable subsets $\{S_i : i \in \Omega\}$, and we define $f(A) = \mu(S_A)$.

The weighted coverage functions can be characterized as the set functions that satisfy the following identities for every pair of disjoint sets R, S with S non-empty:

$$f(R) \geq 0 \tag{4}$$

$$\sum_{T \subseteq S} (-1)^{|T|} f(R \cup T) \leq 0 \tag{5}$$

Submodularity of f is equivalent to the weaker condition that (5) holds for all pairs of disjoint sets R, S such that $|S| = 2$.

Example 7 (Directed cut functions)**.** If $G = (V, E)$ is a directed graph and $A \subseteq V$, let $f(A)$ denote the number of edges $(u, v) \in E$ such that $u \in A$ and $v \notin A$. The easiest way to see that this is a submodular function is to observe that

$$f(A) = g(A) + h(V \setminus A) - |E| \tag{6}$$

where $g(A)$ is the number of edges (u, v) such that $u \in A$, and $h(B)$ is the number of edges (u, v) such that $u \in B$ or $v \in B$. Since g is an additive function and h is a coverage function, equation (6) expresses f as a sum of an additive function, a submodular function, and a constant function, hence f is submodular. (Here, we have applied the principle that if $h : 2^\Omega \to \mathbb{R}$ is a submodular function, then the function $A \mapsto h(\Omega \setminus A)$ is submodular.)

The foregoing arguments apply also to the case in which edges of G have non-negative weights and $f(A)$ is the combined weight of all edges from A to $V \setminus A$. In that case g is still additive and h is a weighted coverage function, so f is still submodular.

Example 8 (Undirected hypergraph cut functions)**.** Let $H = (V, E, w)$ be a finite edge-weighted hypergraph, meaning that V is a finite set, E is a collection of subsets of V (called "hyperedges"), and $w : E \to \mathbb{R}_+$ assigns a non-negative real weight to each set in E. For a set A let $\delta(A)$ denote the set of hyperedges that have at least one vertex in A and at least one vertex in $V \setminus A$. The function $f : 2^V \to \mathbb{R}$ defined by

$$f(A) = \sum_{e \in \delta(A)} w(e)$$

is submodular. In fact,

$$f(A) = h(A) + h(V \setminus A) - |E|$$

where $h(B)$ denotes the sum of the weights of all hyperedges that intersect B, which is a coverage function.

Example 9 (Budgeted additive functions)**.** Given a non-negative valued weight function $w : \Omega \to \mathbb{R}_+$ and a parameter $B \geq 0$, the function

$$f(A) = \min \left\{ \sum_{i \in A} w(i),\ B \right\}$$

is submodular.

Example 10 (Log-determinant functions)**.** If P is a positive definite matrix and Ω denotes the index set of its rows and columns, then for any $A \subseteq \Omega$ we can let P_A denote the square submatrix of P consisting of the entries in the rows and columns indexed by A. The function

$$f(A) = \begin{cases} \log(\det P_A) & \text{if } A \neq \emptyset \\ 1 & \text{if } A = \emptyset \end{cases}$$

is submodular. In fact, since P is positive definite there exists a set of vectors $\{x_i : i \in \Omega\}$ such that $P_{ij} = x_i \cdot x_j$ for all $i, j \in \Omega$. The $|A|$-dimensional volume of the parallelepiped spanned by the vectors $\{x_i : i \in A\}$ is $(\det P_A)^{1/2}$. Using the formula that the volume of a parallelepiped is the area of the base times the height, this implies that

$$(\det P_{A \cup \{i\}})^{1/2} = (\det P_A)^{1/2} \times h_{A,i}^{1/2} \tag{7}$$

where $h_{A,i}$ is the distance from x_i to the linear subspace Λ_A spanned by $\{x_j : j \in A\}$. Taking the logarithm of both sides of (7) we find that

$$f(A \cup \{i\}) - f(A) = \tfrac{1}{2} \log(h_{A,i}). \tag{8}$$

The submodularity of f now follows from the observation that if $A \subseteq B$ then $\Lambda_A \subseteq \Lambda_B$ so the distance of x_i from Λ_A is greater than or equal to its distance from Λ_B.

Example 11 (Entropy functions)**.** For a random variable Y taking values in a set Θ, the Shannon entropy of Y is defined as

$$H(Y) = -\sum_{\theta \in \Theta} \Pr(Y = \theta) \log \Pr(Y = \theta).$$

Let X be a finite probability space (i.e., a probability space with finitely many sample points) and let $\{Y_i : i \in \Omega\}$ be a set of random variables on X taking values in some set Σ. For any subset $A \subseteq \Omega$, let Y_A denote the Σ^A-valued random variable defined by evaluating the tuple $(Y_i)_{i \in A}$. The function

$$f(A) = H(Y_A)$$

is submodular. The proof, which is an application of Jensen's inequalilty for convex functions, can be found in information theory textbooks, for example the textbook by Cover and Thomas.

2 The Lovász Extension

A submodular function f can equivalently be viewed as a function $f : \{0,1\}^\Omega \to \mathbb{R}$ according to the rule that for every $\boldsymbol{x} \in \{0,1\}^\Omega$,

$$f(\boldsymbol{x}) = f(\{i : x_i = 1\}).$$

Our goal in this section is to define a convex function f^- on the domain $[0,1]^\Omega$ that satisfies $f^-(\boldsymbol{x}) = f(\boldsymbol{x})$ for every $x \in \{0,1\}^\Omega$. This will allow us to design an algorithm to minimize submodular functions in polynomial time, by reducing the problem to a convex minimization problem.

In fact, for any function $f : \{0,1\}^\Omega \to \mathbb{R}$, whether or not f is submodular, we may define its *convex closure* to be the function $f^- : [0,1]^\Omega \to \mathbb{R}$ specified by

$$f^-(\boldsymbol{x}) = \min \left\{ \mathbb{E}[f(\boldsymbol{y})] \mid \mathbb{E}[\boldsymbol{y}] = \boldsymbol{x} \right\}, \tag{9}$$

where the minimum is over all probability distributions on vectors $\boldsymbol{y} \in \{0,1\}^\Omega$ that satisfy $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$.

The following easy observations motivate the importance of the convex closure, from an optimization standpoint.

Lemma 1. *For every $f : \{0,1\}^\Omega \to \mathbb{R}$, the convex closure f^- is a convex function on $[0,1]^\Omega$ that satisfies*

$$\forall y \in \{0,1\}^\Omega \quad f^-(\boldsymbol{y}) = f(\boldsymbol{y}) \tag{10}$$

$$\min\{f^-(\boldsymbol{x}) \mid x \in [0,1]^\Omega\} = \min\{f(\boldsymbol{y}) \mid y \in \{0,1\}^\Omega\} \tag{11}$$

$$\max\{f^-(\boldsymbol{x}) \mid x \in [0,1]^\Omega\} = \max\{f(\boldsymbol{y}) \mid y \in \{0,1\}^\Omega\}. \tag{12}$$

Proof. The function f^- is convex because if $\boldsymbol{x}, \boldsymbol{x'}$ are any two vectors in $[0,1]^\Omega$, and p, p' are two distributions on $\{0,1\}^\Omega$ such that

$$\mathbb{E}_{\boldsymbol{y} \sim p}[\boldsymbol{y}] = \boldsymbol{x}, \quad \mathbb{E}_{\boldsymbol{y} \sim p}[f(\boldsymbol{y})] = f^-(\boldsymbol{x})$$
$$\mathbb{E}_{\boldsymbol{y} \sim p'}[\boldsymbol{y}] = \boldsymbol{x'}, \quad \mathbb{E}_{\boldsymbol{y} \sim p'}[f(\boldsymbol{y})] = f^-(\boldsymbol{x'})$$

then for any $\lambda \in [0,1]$ we may define a distribution p'' on $\{0,1\}^\Omega$ by $p''(\boldsymbol{y}) = \lambda p(\boldsymbol{y}) + (1 - \lambda)p'(\boldsymbol{y})$. This satisfies $\mathbb{E}_{\boldsymbol{y} \sim p''}[\boldsymbol{y}] = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x'}$, so

$$f^-(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x'}) \leq \mathbb{E}_{\boldsymbol{y} \sim p''}[f(\boldsymbol{y})] = \lambda f^-(\boldsymbol{x}) + (1-\lambda) f^-(\boldsymbol{x'})$$

which confirms that f^- is convex.

Each $\boldsymbol{y} \in \{0,1\}^\Omega$ is an extreme point of the polyhedron $[0,1]^\Omega$, meaning that the only way to express $\boldsymbol{y}$ as a convex combination $\sum_{\boldsymbol{z}\in\{0,1\}^\Omega} \lambda_{\boldsymbol{z}} \boldsymbol{z}$ is the trivial convex combination in which the coefficient $\lambda_{\boldsymbol{y}}$ equals 1 and all other coefficients are 0. This establishes identity (10). The other two equations are justified by combining (10) with the observation that a weighted average of real numbers in the interval $[a,b]$ always belongs to $[a,b]$. In this case, setting a equal to the right side of (11) and b to the right side of (12), we see that for every $\boldsymbol{x} \in [0,1]^\Omega$ the value $f^-(\boldsymbol{x})$ is a weighted average of numbers in the interval $[a,b]$, so $f^-(\boldsymbol{x})$ itself must belong to $[a,b]$. □

Lemma 1 suggests a method for efficiently computing the minimum of an integer-valued function f on $\{0,1\}^\Omega$, by reducing to minimizing the convex functon f^-. In order for this be a computationally efficient reduction, we must be able to evaluate $f^-(\boldsymbol{x})$ and its subgradient $\nabla_{\boldsymbol{x}}(f^-)$ at any $\boldsymbol{x} \in [0,1]^\Omega$. When f is submodular, there is a very efficient way to evaluate f^- and its subgradient.

Lemma 2. *For any $\boldsymbol{x} \in [0,1]^\Omega$ and $t \in [0,1]$ let*

$$S_{\boldsymbol{x}}(t) = \{j \in \Omega \,:\, \boldsymbol{x}_j > t\}.$$

If f is submodular then its concave closure satisfies

$$\forall \boldsymbol{x} \in [0,1]^\Omega \qquad f^-(\boldsymbol{x}) = \int_0^1 f(S_{\boldsymbol{x}}(t))\, dt. \tag{13}$$

Furthermore the subgradient $\nabla_{\boldsymbol{x}} f^-$ has coordinates given by the formula

$$(\nabla_{\boldsymbol{x}} f^-)_i = f(S_{\boldsymbol{x}}(x_i) \cup \{i\}) - f(S_{\boldsymbol{x}}(x_i)). \tag{14}$$

Proof. Consider any $\boldsymbol{x} \in [0,1]^\Omega$, and suppose p is a distribution on vectors $\boldsymbol{y} \in \{0,1\}^\Omega$ such that $\mathbb{E}_{\boldsymbol{y}\sim p}[\boldsymbol{y}] = \boldsymbol{x}$. The proof will be based upon the observation that among the distributions satisfying this constraint that minimize $\mathbb{E}_{\boldsymbol{y}\sim p}[f(\boldsymbol{y})]$, there is one that is supported on a "nested" set of vectors. Here, we refer to a subset of $\{0,1\}^\Omega$ as "nested" if the vectors in this subset are totally ordered by the $\preceq$ relation, or equivalently if their corresponding sets are totally ordered by the subset relation.

The reason why the set of minimizing distributions should contain one with nested support is that f is submodular. Submodularity of f is relevant because the relations

$$\begin{aligned} \boldsymbol{y}_1 + \boldsymbol{y}_2 &= (\boldsymbol{y}_1 \vee \boldsymbol{y}_2) + (\boldsymbol{y}_1 \wedge \boldsymbol{y}_2) \\ f(\boldsymbol{y}_1) + f(\boldsymbol{y}_2) &\geq f(\boldsymbol{y}_1 \vee \boldsymbol{y}_2) + f(\boldsymbol{y}_1 \wedge \boldsymbol{y}_2) \end{aligned}$$

imply that for any two vectors $\boldsymbol{y}_1, \boldsymbol{y}_2$ in the support of p, if $\delta = \min\{p(\boldsymbol{y}_1), p(\boldsymbol{y}_2)\}$ and q is the distribution

$$q(\boldsymbol{y}) = \begin{cases} p(\boldsymbol{y}) - \delta & \text{if } \boldsymbol{y} \in \{\boldsymbol{y}_1,\, \boldsymbol{y}_2\} \\ p(\boldsymbol{y}) + \delta & \text{if } \boldsymbol{y} \in \{\boldsymbol{y}_1 \vee \boldsymbol{y}_2,\, \boldsymbol{y}_1 \wedge \boldsymbol{2}_2\} \\ p(\boldsymbol{y}) & \text{otherwise.} \end{cases}$$

then $\mathbb{E}_{\boldsymbol{y}\sim q}[\boldsymbol{y}] = \boldsymbol{x}$ and $\mathbb{E}_{\boldsymbol{y}\sim q}[f(\boldsymbol{y})] \leq \mathbb{E}_{\boldsymbol{y}\sim p}[f(\boldsymbol{y})]$. Furthermore, defining Q to be the quadratic function $Q(\boldsymbol{y}) = (\sum_{i\in\Omega} y_i)^2$, if $\boldsymbol{y}_1 \not\preceq \boldsymbol{y}_2$ and $\boldsymbol{y}_2 \not\preceq \boldsymbol{y}_1$ then we have

$$Q(\boldsymbol{y}_1) + Q(\boldsymbol{y}_2) < Q(\boldsymbol{y}_1 \vee \boldsymbol{y}_2) + Q(\boldsymbol{y}_1 \wedge \boldsymbol{y}_2)$$

so

$$\mathbb{E}_{\boldsymbol{y}\sim p}[Q(\boldsymbol{y})] < \mathbb{E}_{\boldsymbol{y}\sim q}[Q(\boldsymbol{y})]$$

Consequently, among all distributions p that satisfy $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{x}$ and that minimize $\mathbb{E}[f(\boldsymbol{y})]$ subject to this constaint, if we choose one that maximizes $\mathbb{E}[Q(\boldsymbol{y})]$ then it must be the case that the support of p is nested.

There is only one distribution with nested support whose expectation is $\boldsymbol{x}$. It is the distribution obtained by sampling $t \in [0, 1]$ uniformly at random and selecting the vector

$$y_i = \begin{cases} 1 & \text{if } x_i > t \\ 0 & \text{otherwise.} \end{cases}$$

The expectation of $f(\boldsymbol{y})$ under this distribution is equal to the right side of (13). The formula (14) for the subgradient of f is then obtained by reasoning about how the right side of (13) varies as we vary $\boldsymbol{x}$ by a small amount. □

3 Packing disjoint arborescences

As one algorithmic application of submodular functions, we present an algorithm for the following network design problem. We are given a directed graph $G = (V, E)$ with a distinguished node r. A *spanning arborescence rooted at* r is defined to be an edge set B such that

1. every vertex $v \neq r$ is the head of exactly one edge in B;
2. r is not the head of any edge in B;
3. for every vertex $v \neq r$, B contains a path from r to v.

In other words, an arborescence is like a directed version of a spanning tree: its underlying set of undirected edges forms a spanning tree of G, and all of its edges are oriented so that they point away from r.

When does a directed graph G contain k edge-disjoint spanning arborescences rooted at r? An obvious necessary condition is that for every vertex $v \neq r$, there are k edge-disjoint paths from r to v. Surprisingly, Edmonds proved that this condition is both necessary and sufficient. Furthermore, there is a polynomial-time algorithm to compute a set of k edge-disjoint spanning arborescences rooted at r, when such a set exists.

Theorem 3 (Edmonds' Branching Theorem)**.** *Suppose G is a directed graph containing at least k edge-disjoint paths from r to v, for every vertex $v \neq r$. Then G contains k edge-disjoint spanning arborescences rooted at r.*

Proof. The following proof is due to László Lovász.

For a vertex set S, let $\delta^G(S)$ denote the set of edges from S to $V \setminus S$, and let $c^G(S)$ be the cut function

$$c^G(S) = |\delta^G(S)|.$$

By Menger's Theorem, the condition that G contains at least k edge-disjoint paths from r to v, for every vertex $v \neq r$, is equivalent to

$$\min\{c^G(S) : \{r\} \subseteq S \subsetneq V\} \geq k. \tag{15}$$

Let $\mathscr{C}$ denote the collection of all vertex sets S such that $\{r\} \subseteq S \subsetneq V$. Assume without loss of generality that the left and right sides of (15) are equal, i.e. k is exactly the minimum value of $c^G(S)$, over all $S \subsetneq V$ that contain r. To prove the theorem, we will show how to construct a spanning arborescence A, rooted at r, such that $c^{G\setminus A}(S) \geq k-1$ for all $S \in \mathscr{C}$. If we present a polynomial time algorithm to construct such an A, it follows that we can iterate the algorithm k times to find k edge-disjoint spanning arborescences rooted at A.

The algorithm to construct A will be a greedy algorithm that initializes $A = \emptyset$ and grows A one edge at a time. Let

$$V(A) = \{r\} \cup \{\text{head}(e) \mid e \in A\}.$$

While $V(A) \neq V$, we will find one edge from $V(A)$ to its complement that can be added to the set A while preserving the constraint that $c^{G\setminus A}(S) \geq k-1$ for all $S \in \mathscr{C}$. Assume (as an inductive hypothesis) that our current set A satisfies this constraint. If $V(A) = V$ then A is a spanning arborescence rooted at r and we are done. Otherwise, there must be an edge $e = (u, v)$ from $V(A)$ to its complement. If we enlarge A to $A \cup \{e\}$ and $V(A)$ to $V(A) \cup \{v\}$ what could go wrong? There could be a vertex set $S \in \mathscr{C}$ such that $c^{G\setminus(A\cup\{e\})}(S) < k-1$. However, this can only happen if $u \in S, v \notin S$, and $c^{G\setminus A}(S) = k-1$. Note that in this case, $v \notin S \cup V(A)$, hence $S \cup V(A) \neq V$. Call a vertex set S *critical with respect to* A if it satisfies

1. $S \in \mathscr{C}$
2. $c^{G\setminus A}(S) = k-1$
3. $S \cup V(A) \neq V$.

We have seen that if there are no critical sets with respect to A, then we can extend A by inserting an arbitrary edge from $V(A)$ to its complement. Otherwise let X be a maximal critical set. We have $c^{G\setminus A}(X \cup V(A)) = c^G(X \cup V(A)) = k$ whereas $c^{G\setminus A}(X) = k-1$, so there is at least one edge $e = (u, v)$ with $u \in V(A) \setminus X$ and $v \notin V(A) \cup X$. Let $A' = A \cup \{e\}$. If $c^{G\setminus A'}(Y) < k-1$ for some $Y \in \mathscr{C}$ then Y must be a critical set with $u \in Y, v \notin Y$. This implies that $Y \cup X \cup V(A) \neq V$ since $v \notin Y \cup X \cup V(A)$. Furthermore the inequalities

$$2(k-1) = c^{G\setminus A}(Y) + c^{G\setminus A}(X) \geq c^{G\setminus A}(Y \cup X) + c^{G\setminus A}(Y \cap X) \geq 2(k-1)$$

imply that $c^{G\setminus A}(Y \cup X) = c^{G\setminus A}(Y \cap X) = k-1$. Hence $Y \cup X$ is a critical set and a strict superset of X, contradicting the maximality of X and hence confirming that $c^{G\setminus A'}(Y) \geq k-1$ for all $Y \in \mathscr{C}$. □

4 Cardinality constrained monotone submodular function maximization

Suppose $f : 2^\Omega \to \mathbb{R}$ is a submodular function that is non-negative and monotone, meaning that $f(A) \geq f(B) \geq 0$ for all $A \supseteq B$. In this section we consider the problem of maximizing the value $f(A)$ over all sets A of some specified cardinality, k. When f is a coverage function defined by a collection of finite sets $\{S_i : i \in \Omega\}$ then the set cover problem asks, for a given k, whether there exists a set $A \subseteq \Omega$ such that $f(A) = |\bigcup_{i\in\Omega} S_i|$. Since set cover is NP-complete, we do not expect a computationally efficient procedure for answering this question. Instead, this section presents and analyzes a natural greedy algorithm for approximately maximizing $f(A)$ over sets A of cardinality k.

The greedy algorithm works as follows: if initializes $A_0 = \emptyset$ and for $\ell = 1, 2, \ldots, k$ it selects A_ℓ to be the set that maximizes $f(A)$ subject to the constraint $|A \setminus A_{\ell-1}| = 1$. In other words, it iteratively adds elements into the set A one by one, each time choosing the element that leads to the greatest marginal increase in the value of f.

To analyze the greedy algorithm let A^* be a set of cardinality k that maximizes $f(A^*)$, and define Δ_ℓ for $\ell = 0, 1, \ldots, k$ by

$$\Delta_\ell = f(A^*) - f(A_\ell).$$

We claim, as an inductive hypothesis, that $\Delta_\ell \leq (1 - \frac{1}{k})^\ell f(A^*)$ for $\ell = 0, 1, \ldots, k$. We have already shown the Since $A_0 = \emptyset$ and $f(\emptyset) \geq 0$, we have $\Delta_0 \leq f(A^*)$, which is the base case of the induction. For the induction step, let us number the elements of A^* as $\{a_1, a_2, \ldots, a_k\}$ and for $j = 0, 1, \ldots, k$ define $B_{\ell,j} = A_\ell \cup \{a_i : i \leq j$. Since $A^* \subseteq A_\ell \cup A^* = B_{\ell,k}$ we have

$$\begin{aligned}
f(A^*) \leq f(B_{\ell,k}) &= f(A_\ell) + \sum_{j=1}^{k} [f(B_{\ell,j}) - f(B_{\ell,j-1})] \\
\Delta_\ell &= \sum_{j=1}^{k} [f(B_{\ell,j}) - f(B_{\ell,j-1})] \\
&\leq \sum_{j=1}^{k} [f(A_\ell \cup \{a_j\}) - f(A_\ell)] \\
\frac{1}{k}\Delta_\ell &\leq \max_{1\leq j\leq k} \{f(A_\ell \cup \{a_j\}) - f(A_\ell) \\
&= f(A_{\ell+1}) - f(A_\ell) \\
\Delta_{\ell+1} &= f(A^*) - f(A_{\ell+1}) = \Delta_\ell - [f(A_{\ell+1}) - f(A_\ell)] \\
&\leq \left(1 - \tfrac{1}{k}\right) \Delta_\ell
\end{aligned}$$

which completes the induction.

Observing that $(1 - \frac{1}{k})^k < \left(e^{-1/k}\right)^k < \frac{1}{e}$, we have proven the following.

Theorem 4 (Nemhauser-Wolsey-Fisher)**.** *The greedy algorithm for monotone non-negative submodular function maximization over sets of cardinality k outputs a solution whose value is at least $1 - (1 - \frac{1}{k})^k > (1 - \frac{1}{e})$ times that of the optimum solution.*

As a corollary, we have the following theorem about the greedy set cover algorithm, which attempts to find a set cover of minimum cardinality by repeatedly choosing the set which covers the greatest number of remaining elements.

Corollary 5. *For a set cover instance specified by sets $S_1, \ldots, S_m$ whose union has cardinality n, the greedy algorithm outputs a set cover whose size is at most $\lceil \ln(n) \rceil$ times greater than the minimum set cover.*

Proof. Suppose the minimum set cover has cardinality k. After the first k iterations of the greedy set cover algorithm, the number of remaining uncovered elements is at most n/e, by Theorem 4. After the next k iterations, the number of remaning uncovered elements is at most n/e^2, by another application of Theorem 4. More generally, after $k \cdot \ell$ iterations of the greedy set cover, the number of remaining uncovered elements is at most n/e^ℓ. When $\ell = \lceil \ln(n) \rceil$ this number is less than 1, so the greedy set cover algorithm must terminate after choosing no more than $k \cdot \lceil \ln(n) \rceil$ sets. □