2 Greedy Strategy

Someone reminded me that I once said, "Greed is good."
Now it seems that it's legal.

— Gordon Gekko (in *Wall Street: Money Never Sleeps*)

I think greed is healthy. You can be greedy
and still feel good about yourself.

— Ivan Boesky

The greedy strategy is a simple and popular idea in the design of approximation algorithms. In this chapter, we study two general theories, based on the notions of independent systems and submodular potential functions, about the analysis of greedy algorithms, and present a number of applications of these methods.

2.1 Independent Systems

The basic idea of a greedy algorithm can be summarized as follows:

(1) We define an appropriate *potential function* $f(A)$ on potential solution sets A.

(2) Starting with $A = \emptyset$, we grow the solution set A by adding to it, at each stage, an element that maximizes (or, minimizes) the value of $f(A \cup \{x\})$, until $f(A)$ reaches the maximum (or, respectively, minimum) value.

D.-Z. Du et al., *Design and Analysis of Approximation Algorithms*,
Springer Optimization and Its Applications 62, DOI 10.1007/978-1-4614-1701-9_2,

We first consider a simple setting, in which the potential function is the same as the objective function. In the following, we write $\mathbb{N}^+$ to denote the set of positive integers, and $\mathbb{R}^+$ the set of nonnegative real numbers.

Let E be a finite set and $\mathcal{I}$ a family of subsets of E. The pair $(E, \mathcal{I})$ is called an *independent system* if

(I_1) $I \in \mathcal{I}$ and $I' \subseteq I \Rightarrow I' \in \mathcal{I}$.

Each subset in $\mathcal{I}$ is called an *independent subset*. Let $c : E \to \mathbb{R}^+$ be a nonnegative function. For every subset F of E, define $c(F) = \sum_{e \in F} c(e)$. Consider the following problem:

> MAXIMUM INDEPENDENT SUBSET (MAX-ISS): Given an independent system $(E, \mathcal{I})$ and a cost function $c : E \to \mathbb{R}^+$,
>
> $$\begin{array}{ll} \text{maximize} & c(I) \\ \text{subject to} & I \in \mathcal{I}. \end{array}$$

We remark that the family $\mathcal{I}$ has, in general, an exponential size and cannot be given explicitly (and, hence, an exhaustive search for the maximum $c(I)$ is impractical). In most applications, however, the system $(E, \mathcal{I})$ is given in such a way that the condition of whether $I \in \mathcal{I}$ can be determined in polynomial time. Under this assumption, the following greedy algorithm, which uses the objective function c as the potential function, works in polynomial time.

Algorithm 2.A (*Greedy Algorithm for* MAX-ISS)

Input: An independent system $(E, \mathcal{I})$ and a cost function $c : E \to \mathbb{R}^+$.

(1) Sort all elements in $E = \{e_1, e_2, \ldots, e_n\}$ in the decreasing order of c. Without loss of generality, assume that $c(e_1) \geq c(e_2) \geq \cdots \geq c(e_n)$.

(2) Set $I \leftarrow \emptyset$.

(3) **For** $i \leftarrow 1$ **to** n **do**
  if $I \cup \{e_i\} \in \mathcal{I}$ **then** $I \leftarrow I \cup \{e_i\}$.

(4) Output $I_G \leftarrow I$. ■

For any instance $(E, \mathcal{I}, c)$ of the problem MAX-ISS, let I^* be its optimal solution and I_G the independent set produced by Algorithm 2.A. We will see that $c(I_G)/c(I^*)$ has a simple upper bound that is independent of the cost function c.

For any $F \subseteq E$, a set $I \subseteq F$ is called a *maximal independent subset* of F if no independent subset of F contains I as a proper subset. For any set $I \subseteq E$, let $|I|$ denote the number of elements in I. Define

$$\begin{aligned} u(F) &= \min\{|I| \mid I \text{ is a maximal independent subset of } F\}, \\ v(F) &= \max\{|I| \mid I \text{ is an independent subset of } F\}. \end{aligned} \tag{2.1}$$

Theorem 2.1 *The following inequality holds for any independent system* $(E, \mathcal{I})$ *and any function* $c : E \to \mathbb{R}^+$*:*

$$1 \leq \frac{c(I^*)}{c(I_G)} \leq \max_{F \subseteq E} \frac{v(F)}{u(F)}.$$

Proof. Assume that $E = \{e_1, e_2, \ldots, e_n\}$, and $c(e_1) \geq \cdots \geq c(e_n)$. Denote $E_i = \{e_1, \ldots, e_i\}$. We claim that $E_i \cap I_G$ is a maximal independent subset of E_i. To see this, we assume, by way of contradiction, that this is not the case; that is, there exists an element $e_j \in E_i \setminus I_G$ such that $(E_i \cap I_G) \cup \{e_j\}$ is independent. Now, consider the jth iteration of the loop of step (3) of Algorithm 2.A. The set I at the beginning of the jth iteration is a subset of I_G, and so $I \cup \{e_j\}$ must be a subset of $(E_i \cap I_G) \cup \{e_j\}$ and, hence, is an independent set. Therefore, the algorithm should have added e_j to I in the jth iteration. This contradicts the assumption that $e_j \notin I_G$.

From the above claim, we see that

$$|E_i \cap I_G| \geq u(E_i).$$

Moreover, since $E_i \cap I^*$ is independent, we have

$$|E_i \cap I^*| \leq v(E_i).$$

Now, we express $c(I_G)$ and $c(I^*)$ in terms of $|E_i \cap I_G|$ and $|E_i \cap I^*|$, respectively. We note that for each $i = 1, 2, \ldots, n$,

$$|E_i \cap I_G| - |E_{i-1} \cap I_G| = \begin{cases} 1, & \text{if } e_i \in I_G, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} c(I_G) &= \sum_{e_i \in I_G} c(e_i) = c(e_1) \cdot |E_1 \cap I_G| + \sum_{i=2}^{n} c(e_i) \cdot (|E_i \cap I_G| - |E_{i-1} \cap I_G|) \\ &= \sum_{i=1}^{n-1} |E_i \cap I_G| \cdot (c(e_i) - c(e_{i+1})) + |E_n \cap I_G| \cdot c(e_n). \end{aligned}$$

Similarly,

$$c(I^*) = \sum_{i=1}^{n-1} |E_i \cap I^*| \cdot (c(e_i) - c(e_{i+1})) + |E_n \cap I^*| \cdot c(e_n).$$

Denote $\rho = \max_{F \subseteq E} v(F)/u(F)$. Then we have

$$\begin{aligned} c(I^*) &\leq \sum_{i=1}^{n-1} v(E_i) \cdot (c(e_i) - c(e_{i+1})) + v(E_n) \cdot c(e_n) \\ &\leq \sum_{i=1}^{n-1} \rho \cdot u(E_i) \cdot (c(e_i) - c(e_{i+1})) + \rho \cdot u(E_n) \cdot c(e_n) \leq \rho \cdot c(I_G). \quad \square \end{aligned}$$

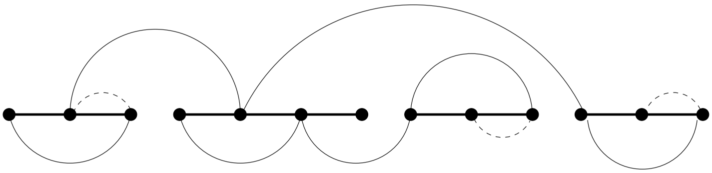

Figure 2.1: Two maximal independent subsets I and J for the problem MAX-HC (the thick lines indicate edges of I, the thin curves and dotted curves indicate the edges of J, and the dotted curves indicate edges shared by I and J).

We note that the ratio $\rho = \max_{F \subseteq E} v(F)/u(F)$ depends only on the structure of the family $\mathcal{I}$ and is independent of the cost function c. Thus, this upper bound is often easy to calculate. We demonstrate the application of this property in two examples.

First, consider the problem MAX-HC defined in Section 1.5. Each instance of this problem consists of n vertices and a distance table on these n vertices. The problem is to find a Hamiltonian circuit of the maximum total distance. Let E be the edge set of the complete graph on the n vertices. Let $\mathcal{I}$ be the family of subsets of E such that $I \in \mathcal{I}$ if and only if I is either a Hamiltonian circuit or a union of disjoint paths (i.e., paths that do not share any common vertex). Clearly, $(E, \mathcal{I})$ is an independent system and whether or not I is in $\mathcal{I}$ can be determined in polynomial time. That is, the problem MAX-HC is a special case of the problem MAX-ISS, and Algorithm 2.A runs on MAX-HC in polynomial time.

Lemma 2.2 *Let $(E, \mathcal{I})$ be the independent system defined above, and F a subset of E. Suppose that I and J are two maximal independent subsets of F. Then $|J| \leq 2|I|$.*

Proof. For $i = 1, 2$, let V_i denote the set of vertices of degree i in I. That is, V_1 is the set of end vertices in I and V_2 is the set of intermediate vertices in I. Clearly, $|I| = |V_2| + |V_1|/2$. Since I is a maximal independent subset of F, every edge in F either is incident on a vertex in V_2 or connects two endpoints of a path in I. Let J_2 be the set of edges in J incident on a vertex in V_2, and $J_1 = J \setminus J_2$. Since J is an independent set, at most two edges in J_2 could be incident on each vertex in V_2. That is, $|J_2| \leq 2|V_2|$. Moreover, every edge in J_1 must connect two endpoints in V_1 in a path of I, and at most one edge in J_1 could be incident on each vertex in V_1. Therefore, $|J_1| \leq |V_1|/2$. (Figure 2.1 shows an example of maximal independent subsets I and J.) Together, we have

$$|J| = |J_1| + |J_2| \leq \frac{|V_1|}{2} + 2|V_2| \leq 2|I|. \qquad \square$$

Theorem 2.3 *When it is applied to the problem* MAX-HC, Algorithm 2.A *is a polynomial-time* 2*-approximation.*

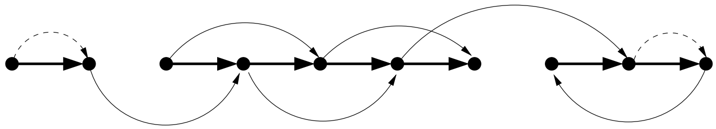

Figure 2.2: Two maximal independent subsets I and J for the problem MAX-DHP.

A similar application gives us a rather weaker performance ratio for the problem MAX-DHP, also defined in Section 1.5. An instance of this problem consists of n vertices and a directed distance table on these n vertices. The problem is to find a directed Hamiltonian path of the maximum total distance. Let E be the set of edges of the complete directed graph on the n vertices. Let $\mathcal{I}$ be the family of subsets of E such that $I \in \mathcal{I}$ if and only if I is a union of disjoint paths. Clearly, $(E, \mathcal{I})$ is an independent system, and whether or not I is in $\mathcal{I}$ can be determined in polynomial time.

Lemma 2.4 *Let $(E, \mathcal{I})$ be the independent system defined as above, and F a subset of E. Suppose that I and J are two maximal independent subsets of F. Then $|J| \leq 3|I|$.*

Proof. Since I is a maximal independent subset of F, every edge in F must have one of the following properties:

(1) It shares a head with an edge in I;

(2) It shares a tail with an edge in I; or

(3) It connects from the head to the tail of a maximal path in I.

(Figure 2.2 shows an example of two maximal independent subsets I and J.)

Let J_1, J_2, and J_3 be the subsets of edges in J that have properties (1), (2) and (3), respectively. Since J is an independent subset, each edge in I can share its head (or its tail) with at most one edge in J, and each maximal path in I can be connected from the head to the tail by at most one edge in J. That is, $|J_i| \leq |I|$, for $i = 1, 2, 3$. Thus,

$$|J| = |J_1| + |J_2| + |J_3| \leq 3|I|. \qquad \square$$

Theorem 2.5 *When it is applied to the problem* MAX-DHP*,* Algorithm 2.A *is a polynomial-time* 3*-approximation.*

The following simple example shows that the performance ratio given by the above theorem cannot be improved.

Example 2.6 Consider the following distance table on four vertices, in which the parameter ε is a positive real number less than 1:

	a	b	c	d
a	0	1	ε	ε
b	ε	0	1	ε
c	ε	$1+\varepsilon$	0	1
d	ε	ε	ε	0

It is clear that the longest Hamiltonian path has distance 3 and yet the greedy algorithm selects the edge (c, b) first and gets a path of total distance $1 + 3\varepsilon$. The performance ratio is, thus, equal to $3/(1+3\varepsilon)$, which approaches 3 when ε approaches zero. $\square$

2.2 Matroids

Let E be a finite set and $\mathcal{I}$ a family of subsets of E. The pair $(E, \mathcal{I})$ is called a *matroid* if

(I_1) $I \in \mathcal{I}$ and $I' \subseteq I \Rightarrow I' \in \mathcal{I}$; and

(I_2) For any subset F of E, $u(F) = v(F)$,

where $u(F)$ and $v(F)$ are the two functions defined in (2.1). Thus, an independent system $(E, \mathcal{I})$ is a matroid if and only if, for any subset F of E, all maximal independent subsets of F have the same cardinality. From Theorem 2.1, we know that Algorithm 2.A produces an optimal solution for the problem MAX-ISS if the input instance $(E, \mathcal{I})$ is a matroid. The next theorem shows that this property actually characterizes the notion of matroids.

Theorem 2.7 *An independent system* $(E, \mathcal{I})$ *is a matroid if and only if for every nonnegative function* $c : E \to \mathbb{R}^+$*, the greedy* Algorithm 2.A *produces an optimal solution for the instance* $(E, \mathcal{I}, c)$ *of* MAX-ISS.

Proof. The "only if" part is just Theorem 2.1. Now, we prove the "if" part. Suppose that $(E, \mathcal{I})$ is not a matroid. Then we can find a subset F of E such that F has two maximal independent subsets I and I' with $|I| > |I'|$. Define, for any $e \in E$,

$$c(e) = \begin{cases} 1+\epsilon, & \text{if } e \in I', \\ 1, & \text{if } e \in I \setminus I', \\ 0, & \text{if } e \in E \setminus (I \cup I'), \end{cases}$$

where ϵ is a positive number less than $1/|I'|$ (so that $c(I) > c(I')$). Clearly, for this cost function c, Algorithm 2.A produces the solution set I', which is not optimal. $\square$

The following are some examples of matroids.

Example 2.8 Let E be a finite set of vectors and $\mathcal{I}$ the family of linearly independent subsets of E. Then the size of the maximal independent subset of a subset $F \subseteq E$ is the rank of F and is unique. Thus, $(E, \mathcal{I})$ is a matroid. $\square$

Example 2.9 Given a graph $G = (V, E)$, let $\mathcal{I}$ be the family of edge sets of acyclic subgraphs of G. Then it is clear that $(E, \mathcal{I})$ is an independent system. We verify that it is actually a matroid, which is usually called a *graph matroid*.

Consider a subset F of E. Suppose that the subgraph (V, F) of G has m connected components. We note that in each connected component C of (V, F), a maximal acyclic subgraph is just a spanning tree of C, in which the number of edges is exactly one less than the number of vertices in C. Thus, every maximal acyclic subgraph of (V, F) has exactly $|V| - m$ edges. So, condition (I_2) holds for the independent system $(E, \mathcal{I})$, and hence $(E, \mathcal{I})$ is a matroid. □

Example 2.10 Consider a directed graph $G = (V, E)$ and a nonnegative integer function f on V. Let $\mathcal{I}$ be the family of edge sets of subgraphs whose out-degree at any vertex u is no more than $f(u)$. It is clear that $(E, \mathcal{I})$ is an independent system. We verify that $(E, \mathcal{I})$ is actually a matroid.

For any subset $F \subseteq E$, let $d_F^+(u)$ be the number of out-edges at u which belong to F. Then, all maximal independent sets in F have the same size,

$$\sum_{u \in V} \min\{f(u), d_F^+(u)\}.$$

Therefore, $(E, \mathcal{I})$ is a matroid. □

In a matroid, all maximal independent subsets have the same cardinality. They are called *bases*. For instance, in a graph matroid defined by a connected graph $G = (V, E)$, every base is a spanning tree of G and they all have the same size $|V| - 1$.

There is an interesting relationship between the intersection of matroids and independent systems.

Theorem 2.11 *For any independent system $(E, \mathcal{I})$, there exist a finite number of matroids $(E, \mathcal{I}_i)$, $1 \leq i \leq k$, such that $\mathcal{I} = \bigcap_{i=1}^{k} \mathcal{I}_i$.*

Proof. Let $C_1, \ldots, C_k$ be all minimal dependent sets of $(E, \mathcal{I})$ (i.e, they are the minimal sets among $\{F \mid F \subseteq E, F \notin \mathcal{I}\}$). For each $i \in \{1, 2, \ldots, k\}$, define

$$\mathcal{I}_i = \{F \subseteq E \mid C_i \not\subseteq F\}.$$

Then it is not hard to verify that $\mathcal{I} = \bigcap_{i=1}^{k} \mathcal{I}_i$. We next show that each $(E, \mathcal{I}_i)$ is a matroid.

It is easy to see that $(E, \mathcal{I}_i)$ is an independent system. Thus, it suffices to show that condition (I_2) holds for $(E, \mathcal{I}_i)$. Consider $F \subseteq E$. If $C_i \not\subseteq F$, then F contains a unique maximal independent set, which is itself. If $C_i \subseteq F$, then every maximal independent subset of F is equal to $F \setminus \{u\}$ for some $u \in C_i$ and hence has size $|F| - 1$. □

Theorem 2.12 *Suppose the independent system $(E, \mathcal{I})$ is the intersection of k matroids $(E, \mathcal{I}_i)$, $1 \leq i \leq k$; that is, $\mathcal{I} = \bigcap_{i=1}^{k} \mathcal{I}_i$. Then*

$$\max_{F\subseteq E} \frac{v(F)}{u(F)} \le k,$$

where $u(F)$ *and* $v(F)$ *are the two functions defined in* (2.1).

Proof. Let $F \subseteq E$. Consider two maximal independent subsets I and J of F with respect to $(E, \mathcal{I})$. For each $1 \le i \le k$, let I_i be a maximal independent subset of $I \cup J$ with respect to $(E, \mathcal{I}_i)$ that contains I. [Note that I is an independent subset of $I \cup J$ with respect to $(E, \mathcal{I}_i)$, and so such a set I_i exists.] For any $e \in J \setminus I$, if $e \in \bigcap_{i=1}^k (I_i \setminus I)$, then $I \cup \{e\} \in \bigcap_{i=1}^k \mathcal{I}_i = \mathcal{I}$, contradicting the maximality of I. Hence, e occurs in at most $k-1$ different subsets $I_i \setminus I$. It follows that

$$\sum_{i=1}^{k} |I_i| - k|I| = \sum_{i=1}^{k} |I_i \setminus I| \le (k-1)|J \setminus I| \le (k-1)|J|,$$

or

$$\sum_{i=1}^{k} |I_i| \le k|I| + (k-1)|J|.$$

Now, for each $1 \le i \le k$, let J_i be a maximal independent subset of $I \cup J$ with respect to $(E, \mathcal{I}_i)$ that contains J. Since, for each $1 \le i \le k$, $(E, \mathcal{I}_i)$ is a matroid, we must have $|I_i| = |J_i|$. In addition, for every $1 \le i \le k$, $|J| \le |J_i|$. Therefore, we get

$$k|J| \le \sum_{i=1}^{k} |J_i| = \sum_{i=1}^{k} |I_i| \le k|I| + (k-1)|J|.$$

It follows that $|J| \le k|I|$. □

Example 2.13 Consider the independent system $(E, \mathcal{I})$ for MAX-DHP defined in Section 2.1. Based on the analysis in the proof of Lemma 2.4 and Examples 2.9 and 2.10, we can see that $\mathcal{I}$ is actually the intersection of the following three matroids:

(1) The family $\mathcal{I}_1$ of all subgraphs with out-degree at most 1 at each vertex;

(2) The family $\mathcal{I}_2$ of all subgraphs with in-degree at most 1 at each vertex; and

(3) The family $\mathcal{I}_3$ of all subgraphs that do not contain a cycle when the edge direction is ignored.

Thus, Theorem 2.5 can also be derived from Theorem 2.12.

On the other hand, for the independent system $(E, \mathcal{I})$ for MAX-HC defined in Section 2.1, the analysis in the proof of Lemma 2.2 uses a more complicated counting argument and does not yield the simple property that $(E, \mathcal{I})$ is the intersection of two matroids. In fact, it can be proved that $(E, \mathcal{I})$ is *not* the intersection of two matroids. We remark that, in general, the problem MAX-ISS for an independent system that is the intersection of two matroids can often be solved in polynomial time. □

Example 2.14 Let X, Y, Z be three sets. We say two elements (x_1, y_1, z_1) and (x_2, y_2, z_2) in $X \times Y \times Z$ are *disjoint* if $x_1 \neq x_2$, $y_1 \neq y_2$, and $z_1 \neq z_2$. Consider the following problem:

> MAXIMUM 3-DIMENSIONAL MATCHING (MAX-3DM): Given three disjoint sets X, Y, Z and a nonnegative weight function c on all triples in $X \times Y \times Z$, find a collection $\mathcal{F}$ of disjoint triples with the maximum total weight.

For given sets X, Y, and Z, let $E = X \times Y \times Z$. Also, let $\mathcal{I}_X$ $(\mathcal{I}_Y, \mathcal{I}_Z)$ be the family of subsets A of E such that no two triples in any subset share an element in X $(Y, Z$, respectively). Then $(E, \mathcal{I}_X)$, $(E, \mathcal{I}_Y)$, and $(E, \mathcal{I}_Z)$ are three matroids and MAX-3DM is just the problem of finding the maximum-weight intersection of these three matroids. By Theorem 2.12, we see that Algorithm 2.A is a polynomial-time 3-approximation for MAX-3DM. □

2.3 Quadrilateral Condition on Cost Functions

Theorem 2.7 gives us a tight relationship between matroids and the optimality of greedy algorithms. It is interesting to point out that this tight relationship holds with respect to *arbitrary* nonnegative objective functions c. That is, if $(E, \mathcal{I})$ is a matroid, then the greedy algorithm will find optimal solutions for all objective functions c. On the other hand, if $(E, \mathcal{I})$ is not a matroid, then the greedy algorithm may still produce an optimal solution, but the optimality must depend on some specific properties of the objective functions. In this section, we present such a property.

Consider a directed graph $G = (V, E)$ and a cost function $c : E \to \mathbb{R}$. We say (G, c) satisfies the *quadrilateral condition* if, for any four vertices u, v, u', v' in V,

$$\begin{aligned} c(u, v) \geq \max\{c(u, v'), c(u', v)\} \\ \Longrightarrow c(u, v) + c(u', v') \geq c(u, v') + c(u', v). \end{aligned}$$

The quadrilateral condition is quite useful in the analysis of greedy algorithms. The following are some examples.

Let $G = (V_1, V_2, E)$ be a complete bipartite graph with $|V_1| = |V_2|$. Let $\mathcal{I}$ be the family of all matchings (recall that a *matching* of a graph is a set of edges that do not share any common vertex). Clearly, $(E, \mathcal{I})$ is an independent system. It is, however, not a matroid. In fact, for some subgraphs of G, maximal matchings may have different cardinalities (although all maximal matchings for G always have the same cardinality). A maximal matching in the bipartite graph is called an *assignment*.

> MAXIMUM ASSIGNMENT (MAX-ASSIGN): Given a complete bipartite graph $G = (V_1, V_2, E)$ with $|V_1| = |V_2|$, and an edge weight function $c : E \to \mathbb{R}^+$, find a maximum-weight assignment.

Theorem 2.15 *If the weight function c satisfies the* quadrilateral condition *for all $u, u' \in V_1$ and $v, v' \in V_2$, then* Algorithm 2.A *produces an optimal solution for the instance (G, c) of* MAX-ASSIGN.