Chapter 2
Maximization of Submodular Functions: Theory and Algorithms

2.1 Introduction

In this chapter we give some theoretical results fundamental to the problem of finding a global maximum of a general submodular (or, equivalently, global minimum of a general supermodular) set function (see [73] which we call the problem of maximization of submodular functions (PMSFs) (following [102]). By a set function we mean a mapping from 2^N to the real numbers, where $N = \{1, 2, \ldots, n\}$. Another well-known term for an arbitrary set function is a *pseudo-Boolean function* (see [84], [20, 21]) which is a mapping from $\{0, 1\}^n$ to the real numbers. PMSF is known to be NP-hard, though the corresponding minimization problem is known to be polynomially solvable (see, e.g., [127]). Enormous interest in studying PMSF arises from the fact that several classes of important combinatorial optimization problems belong to PMSF, including the Simple or "Uncapacitated" Plant (Facility) Location Problem (SPLP) and its competitive version (see [11]), the quadratic cost partition (QCP) problem with nonnegative edge weights, and its special case—the Max-Cut Problem, the generalized transportation problem [109, 110]. Many models in mathematics [103], including the rank function of elementary linear algebra, which is a special case of matroid rank functions (see [42, 45]), require the solution of a PMSF.

Although the general problem of the maximization of a submodular function is known to be NP-hard, there has been a sustained research effort aimed at developing practical procedures for solving medium and large-scale instances of the PMSF. Often the approach taken has been problem specific, and supermodularity of the underlying objective function has been only implicit to the analysis. For example, Barahona et al. [7] have addressed the max-cut problem from the point of view of polyhedral combinatorics and developed a branch and cut algorithm, suitable for applications in statistical physics and circuit layout design. Beasley [9] applies Lagrangean heuristics to several classes of location problems including SPLPs and

B. Goldengorin and P.M. Pardalos, *Data Correcting Approaches in Combinatorial Optimization*, SpringerBriefs in Optimization, DOI 10.1007/978-1-4614-5286-7_2,

reports results of extensive experiments on a Cray supercomputer. Lee et al. [102] have made a study of the QCP problem of which max-cut with nonnegative edge weights is a special case, again from the standpoint of polyhedral combinatorics.

There have been fewer published attempts to develop algorithms for minimization of a general supermodular function. We believe that the earliest attempt to exploit supermodularity is the work of [118], who identified a supermodular structure in their study of railway timetabling. Their procedure was subsequently published by Cherenin [29] as the "method of successive calculations." Their algorithm however is not widely known in the West [4] where, as far as we are aware of, the only general procedures that have been studied in depth are the greedy approximation algorithm from [109], and the algorithm for maximization of submodular functions subject to linear constraints from [110]. In a comment to a note by Frieze [46], Babayev [4] demonstrated that Frieze's two rules: OP1 and OP2, developed to accelerate a BnB algorithm for the SPLP were a consequence of Cherenin's theorem for PMSF [29]. Note that Alcouffe and Muratet's [3] algorithm is based on a special case of Cherenin's [29] "method of successive calculations."

Indeed the only practical algorithmic implementation known in the West appears to be the "accelerated greedy" (AG) algorithm of [105], which has been applied to optimal planning and design of telecommunication networks. We note that the AG algorithm has also been applied to the problem of D-optimal experimental design [125]; see also Ko et al. [92] and Lee [101] for further examples of "hard" D-optimal design problems in environmental monitoring. In [50] an optimal algorithm is constructed with exponential time complexity for the well-known Shannon max–min problem. This algorithm is applied to the maximization of submodular functions subject to a convex set of feasible solutions, and to the problem of—what is known as—decoding monotonic Boolean functions.

In this chapter we present an elegant key theorem of Cherenin, which provides the basis of excluding rules, and in particular, for the justification of the Preliminary Preservation (Dichotomy) algorithm. We generalize Cherenin's excluding rules in the form of "preservations rules" which will be used in Chap. 3. Moreover, our preservation rules can be used for implicit enumeration of subproblems in a BnB approach for solving PMSF.

The chapter is organized as follows. In Sect. 2.2 we motivate a theoretical development of these rules by presenting some important results on the structure of local and global maxima for submodular functions by Cherenin [29] and Khachaturov [89, 90]. In this section a fundamental theorem of Cherenin is stated, which provides the basis of "the method of successive calculations." Section 2.2 also contains an important characterization of local maxima as disjoint components of "strict" and "saddle" vertices which greatly assists the understanding of the difference between the properties of Cherenin's "excluding rules" and our "preservation rules" discussed in Sect. 2.3. In Sect. 2.4 we present our main Theorem 2.8 from which generalized bounds for implicit enumeration can be derived, and allow the rules of Sect. 2.3 to be extended to other cases (ε-*optimality*). We present the two different representations (a) and (b) of the partition of the current set of feasible solutions (vertices) defined by a strictly inner vertex with respect to this set.

By using our main Theorem 2.8 and representations (a) and (b), we prove the correctness of Cherenin's excluding rules in the form of our preservation rules. These rules are the basis of Cherenin's preliminary preservation algorithm (PPA) [118]. We introduce the so-called *nonbinary branching rules*, based on Theorem 2.8 in Sect. 2.6. Nonbinary branching rules are illustrated by an instance of the SPLP. In Sect. 2.5 we outline the main steps of the PPA and illustrate how our new preservation rules (see Corollary 2.6) can be applied to a small example of the SPLP. We show that if the PPA terminates with a global maximum, then the given submodular function has exactly one strict component. Section 2.7 gives a number of concluding remarks.

2.2 The Structure of Local and Global Maxima of Submodular Set Functions

In this section we present results of Cherenin–Khachaturov (see [29, 89]) which are hardly known in the Western literature (see also [4]).

Let z be a real-valued function defined on the power set 2^N of $N = \{1, 2, \ldots, n\}$; $n \geq 1$. For each $S, T \in 2^N$ with $S \subseteq T$, define

$$[S,T] = \{I \in 2^N \mid S \subseteq I \subseteq T\}.$$

Note that $[\emptyset, N] = 2^N$. Any *interval* $[S,T]$ is, in fact, a *subinterval* of $[\emptyset, N]$ if $\emptyset \subseteq S \subseteq T \subseteq N$; notation $[S,T] \subseteq [\emptyset, N]$. In this book we mean by an interval always a subinterval of $[\emptyset, N]$. Throughout this book we consider a set of PMSFs defined on any interval $[S,T] \subseteq [\emptyset, N]$ as follows:

$$\max\{z(I) \mid I \in [S,T]\} = z^*[S,T], \text{ for all } [S,T] \subseteq [\emptyset, N].$$

The function z is called *submodular* on $[S,T]$ if for each $I, J \in [S,T]$ it holds that

$$z(I) + z(J) \geq z(I \cup J) + z(I \cap J).$$

Expressions of the form $S \setminus \{k\}$ and $S \cup \{k\}$ will be abbreviated to $S - k$ and $S + k$.

The following theorem presented in [109] gives a number of equivalent formulations for submodular functions which is useful for a clearer understanding of the concept of submodularity. Since sometime we use the incremental or decremental value of $z(S)$, we define $d_j^+(S) = z(S + j) - z(S)$ and $d_j^-(S) = z(S - j) - z(S)$.

Theorem 2.1. *All the following statements are equivalent and define a submodular function.*

$$(i)\ z(A) + z(B) \geq z(A \cup B) + z(A \cap B),\ \ \forall A, B \subseteq N.$$

$$(ii)\ \ d_j^+(S) \geq d_j^+(T),\ \ \forall S \subseteq T \subseteq N \text{ and } j \in N \setminus T.$$

$$(iii)\ d_j^+(S) \geq d_j^+(S+k),\ \ \forall S \subseteq N \text{ and } j \in N \setminus (S+k) \text{ and } k \in N \setminus S.$$

$$(iv) \qquad z(T) \leq z(S) + \sum_{j \in T \setminus S} d_j^+(S), \quad \forall S \subseteq T \subseteq N.$$

$$(v) \qquad z(S) \leq z(T) + \sum_{j \in T \setminus S} d_j^-(T), \quad \forall S \subseteq T \subseteq N.$$

As an example consider the QCP problem, for which it is well known that the objective function $z(Q)$ is a submodular function (see e.g., [102]). For given real numbers p_i and nonnegative real numbers q_{ij} with $i, j \in N$, the QCP is the problem of finding a subset Q of N such that the weight $z(Q) = \sum_{i \in Q} p_i - \frac{1}{2} \sum_{i,\, j \in Q} q_{ij}$ is as large as possible. Let N be the vertex set, $E \subseteq N \times N$ the edge set of an edge-weighted graph $G = (N, E)$, and $w_{ij} \geq 0$ are edge weights. For each $Q \subseteq N$, the cut $\delta(Q)$ is defined as the edge set for which each edge has one end in Q and the other one in $N \setminus Q$. It is easy to see that the Max-Cut Problem with nonnegative edge weights is a QCP where $p_i = \sum_{j \in N} w_{ij}$ and $q_{ij} = 2w_{ij}$, for $i, j \in N$.

Lemma 2.1. *The objective $z(S)$ of the QCP problem is submodular.*

Proof. According to Theorem 2.1(iii) a function is submodular if

$$d_l^+(S) \geq d_l^+(S+k),\ \ \forall S \subseteq N \text{ and } l \in N \setminus (S+k) \text{ and } k \in N \setminus S.$$

Substituting $d_l^+(S) = z(S+l) - z(S)$ we get

$$z(S+l) - z(S) \geq z(S+k+l) - z(S+k)$$

Substituting $z(S) = \sum_{i \in S} p_i - \frac{1}{2} \sum_{i,j \in S} q_{ij}$ gives

$$\sum_{i \in S+l} p_i - \frac{1}{2} \sum_{i,j \in S+l} q_{ij} - \left(\sum_{i \in S} p_i - \frac{1}{2} \sum_{i,j \in S} q_{ij} \right)$$

$$\geq \sum_{i \in S+k+l} p_i - \frac{1}{2} \sum_{i,j \in S+k+l} q_{ij} - \left(\sum_{i \in S+k} p_i - \frac{1}{2} \sum_{i,j \in S+k} q_{ij} \right)$$

Canceling out terms involving p_i, we obtain

$$- \sum_{i,j \in S+l} q_{ij} + \sum_{i,j \in S} q_{ij} \geq - \sum_{i,j \in S+k+l} q_{ij} + \sum_{i,j \in S+k} q_{ij}$$

This result, after some bookkeeping, implies

$$q_{kl} + q_{lk} \geq 0$$

Since q_{ij} is nonnegative for all $i, j \in N$, the proof is completed. □

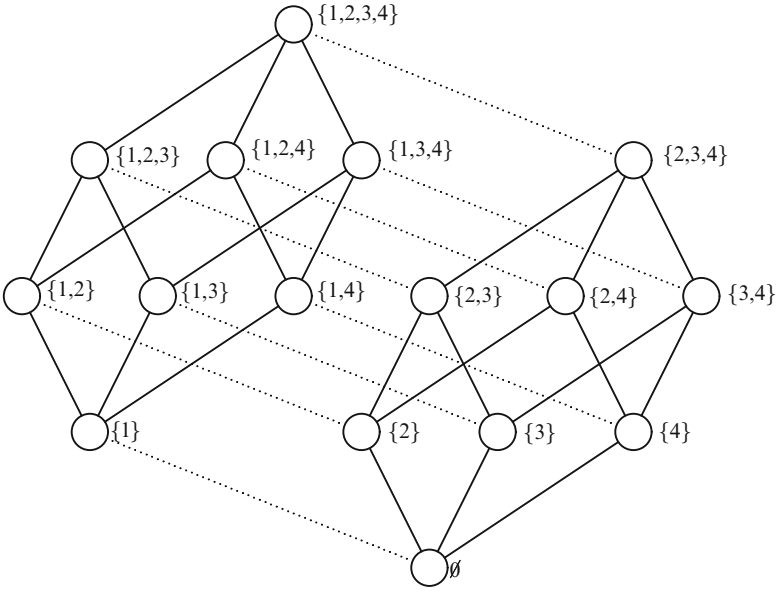

Fig. 2.1 The Hasse diagram of $\{1,2,3,4\}$

Hence, the QCP problem is a special case of the problem of maximizing a submodular function.

A subset $L \in [\emptyset, N]$ is called a *local maximum* of z if for each $i \in N$,

$$z(L) \geq \max\{z(L-i), z(L+i)\}.$$

A subset $S \in [\emptyset, N]$ is called a *global maximum* of z if $z(S) \geq z(I)$ for each $I \in [\emptyset, N]$. We will use the Hasse diagram (see e.g., [76] and Fig. 2.1) as the ground graph $G = (V, E)$ in which $V = [\emptyset, N]$ and a pair (I, J) is an edge if either $I \subset J$ or $J \subset I$, and $|I \setminus J| + |J \setminus I| = 1$.

The graph $G = (V, E)$ is called *z-weighted* if the weight of each vertex $I \in V$ is equal to $z(I)$; notation $G = (V, E, z)$. In terms of $G = (V, E, z)$ the PMSF means finding a vertex $S \in V$ of the weight $z(S)$ which is as large as possible. An example of the weighted G with $N = \{1,2,3,4\}$ is depicted in Fig. 2.2, where the weight $z(I)$ is indicated inside the corresponding vertex I.

Here among others the vertices $\{1,2,3\}$ and $\{4\}$ are local maxima, and $\{4\}$ is a global maximum (see Fig. 2.2).

A sequence $\Gamma = (I^0, I^1, \ldots, I^n)$ of subsets $I^t \in 2^N$, $t = 0, 1, \ldots, n$ such that $|I^t| = t$ and

$$\emptyset = I^0 \subset I^1 \subset I^2 \subset \ldots \subset I^t \subset \ldots \subset I^{n-1} \subset I^n = N$$

is called a *chain* in $[\emptyset, N]$. An example of the chain $\emptyset \subset \{2\} \subset \{2,4\} \subset \{1,2,4\} \subset \{1,2,3,4\}$ in Fig. 2.3 is shown.

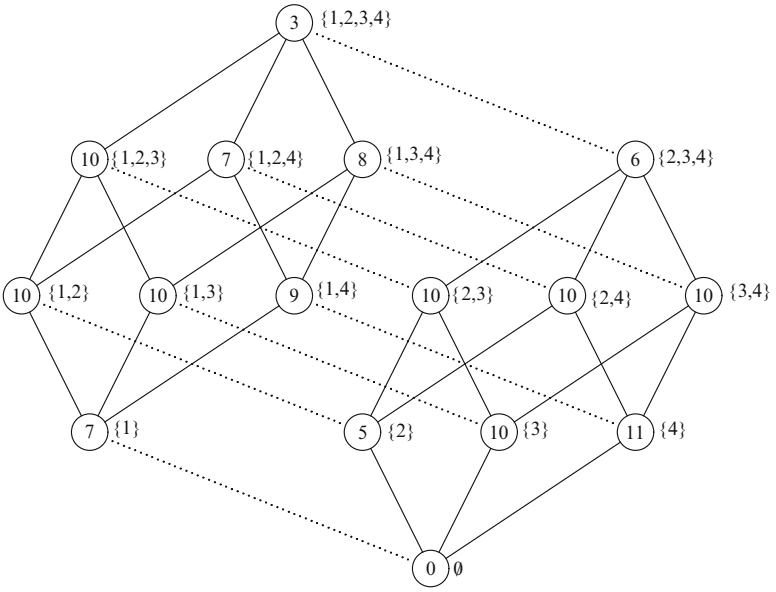

Fig. 2.2 Example of local maxima $\{1,2\}$, $\{1,2,3\}$, $\{1,3\}$, $\{2,3\}$, $\{3\}$, and the global maximum $\{4\}$ on the Hasse diagram

Similarly, a chain of any interval $[S,T]$ can be defined. A submodular function z is *nondecreasing (nonincreasing)* on the chain Γ if $z(I^l) \leq z(I^m)$ ($z(I^l) \geq z(I^m)$) for all l, m such that $0 \leq l \leq m \leq n$; concepts of *increasing, decreasing, and constant* (signs, respectively, $<,>,=$) are defined in an obvious manner (see, for example, Fig. 2.4).

The following theorem (see [29]) shows the quasiconcavity of a submodular function on any chain that includes a local maximum (see Fig. 2.5).

Theorem 2.2. *Let z be a submodular function on 2^N and let L be a local maximum. Then z is nondecreasing on any chain in $[\emptyset,L]$, and nonincreasing on any chain in $[L,N]$.*

Proof. We show that z is nondecreasing on any chain in $[\emptyset,L]$. If either $L=\emptyset$ (we obtain the nonincreasing case) or $|L|=1$, the assertion is true, since L is a local maximum of z. So, let $|L|>1$ and $I,J \in [\emptyset,L]$ such that $J=I+k$, $k \in L \setminus I$.

Note that $\emptyset \subseteq \ldots \subseteq I \subset J \subseteq \ldots \subset L$. The submodularity of z implies $z(J)+z(L-k) \geq z(I)+z(L)$, or $z(J)-z(I) \geq z(L)-z(L-k)$. Since L is a local maximum, $z(L)-z(L-k) \geq 0$. Hence $z(J) \geq z(I)$, and we have finished the proof of the nondecreasing case. The proof for $[L,N]$ is similar. □

Corollary 2.1. *Let z be a submodular function on 2^N and let L_1 and L_2 be local maxima with $L_1 \subseteq L_2$. Then z is a constant on $[L_1,L_2]$, and every $L \in [L_1,L_2]$ is a local maximum of z.*

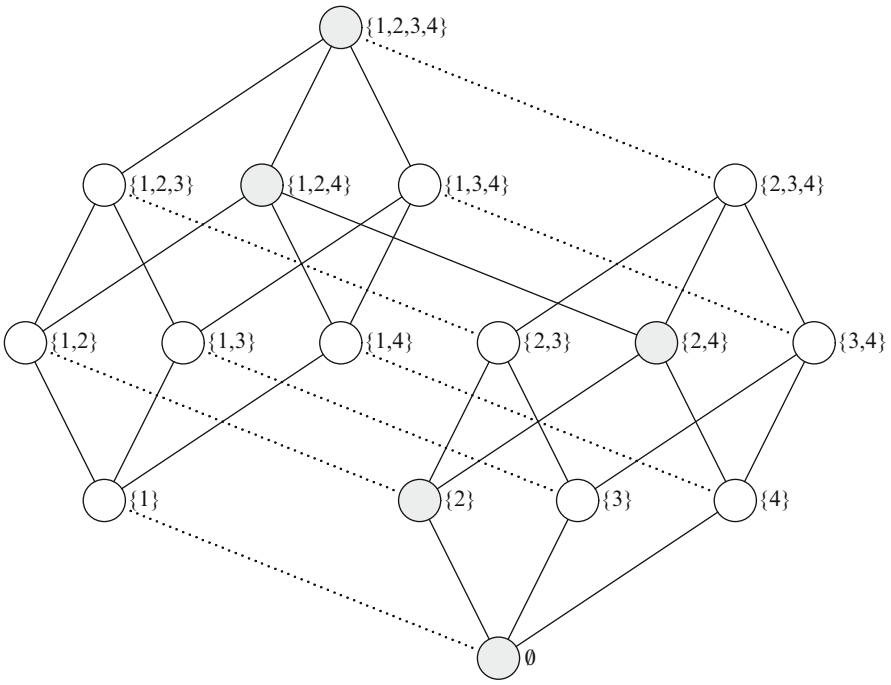

Fig. 2.3 Example of the chain $\emptyset \subset \{2\} \subset \{2,4\} \subset \{1,2,4\} \subset \{1,2,3,4\}$ in the Hasse diagram of $\{1,2,3,4\}$

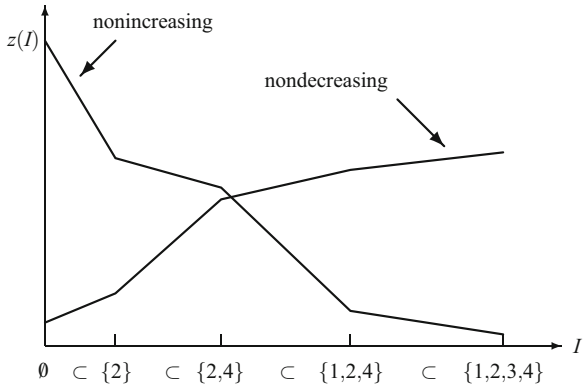

Fig. 2.4 Example of a nondecreasing (nonincreasing) function on the chain in the Hasse diagram of $\{1,2,3,4\}$

Proof. First we show that z is a constant function on $[L_1, L_2]$. Let us apply Theorem 2.2 to a chain including $\emptyset \subseteq \cdots \subseteq L_1 \subseteq L_2 \subseteq \cdots \subseteq N$, first to the single (isolated) local maximum L_2 and second to the single local maximum L_1. For the first case we obtain $z(\emptyset) \leq \cdots \leq z(L_1) \leq \cdots \leq z(I) \leq z(L_2)$. For any subchain of

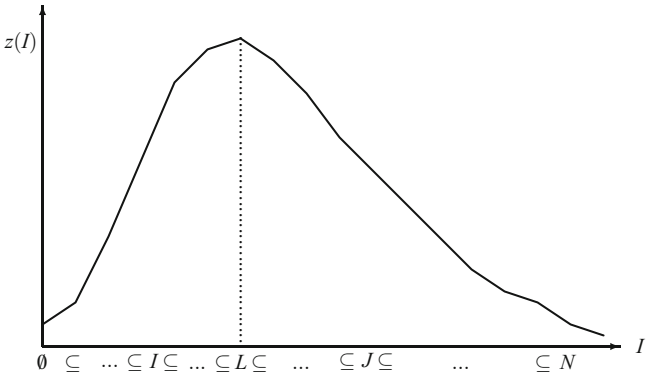

Fig. 2.5 A quasiconcave behavior of a submodular function on the chain with a local maximum L (Cherenin's theorem)

the interval $[L_1, L_2]$ we have $z(L_1) \leq \cdots \leq z(L_2)$. By the same reasoning for the second case we have $z(L_1) \geq \cdots \geq z(L_2)$. Combining both sequences of inequalities we have proved the constancy of z.

Now we show that every $L \in [L_1, L_2]$ is a local maximum of z. Assume the contrary that there exists $L \in [L_1, L_2]$ that is not a local maximum of z. Then either there is a $L - i \notin [L_1, L_2]$ with $z(L) < z(L - i)$ or there is a $L + i \notin [L_1, L_2]$ with $z(L) < z(L + i)$. In the first case we get accordingly the definition of submodularity $z(L) + z(L_2 - i) \geq z(L - i) + z(L_2)$ or $z(L) - z(L - i) \geq z(L_2) - z(L_2 - i) \geq 0$. This contradicts $z(L) < z(L - i)$. In the second case a similar argument holds by using L_1 instead of L_2. □

In Corollary 2.1 we have indicated two important structural properties of a submodular function considered on intervals whose end points are local maxima. Namely, on such an interval a submodular function preserves a constant value and every point of this interval is a local maximum. It will be natural to consider the widest intervals with the above-mentioned properties.

A local maximum $\underline{L} \in 2^N$ ($\overline{L} \in 2^N$) is called a *lower* (respectively, *upper*) *maximum* if there is no other local maximum L such that $L \subset \underline{L}$ (respectively, $\overline{L} \subset L$). For example, in Fig. 2.6 the vertex $\{1,2,3\}$ is an upper local maximum and the vertices $\{1,2\}$, $\{3\}$ are lower local maxima. Note that the vertex $\{3,4\}$ is not a local maximum. If an interval $[\underline{L}, \overline{L}]$ with $\underline{L} \subseteq \overline{L}$ has as its end points lower and upper maxima, then it is the widest interval on which the submodular function is a constant and each point is a local maximum. We call a pair of intervals $[\underline{L}_i, \overline{L}_i]$ with $\underline{L}_i \subseteq \overline{L}_i$, $i = 1,2$ *connected* if $[\underline{L}_1, \overline{L}_1] \cap [\underline{L}_2, \overline{L}_2] \neq \emptyset$. The intervals of local maxima form a set of *components of local maxima*. Two intervals belong to the same component if they are connected. Hence, two local maxima L_1 and L_2 are in the same component if there is a path in $G = (V, E, z)$ with end vertices L_1 and L_2, and all intermediate vertices of this path are local maxima.

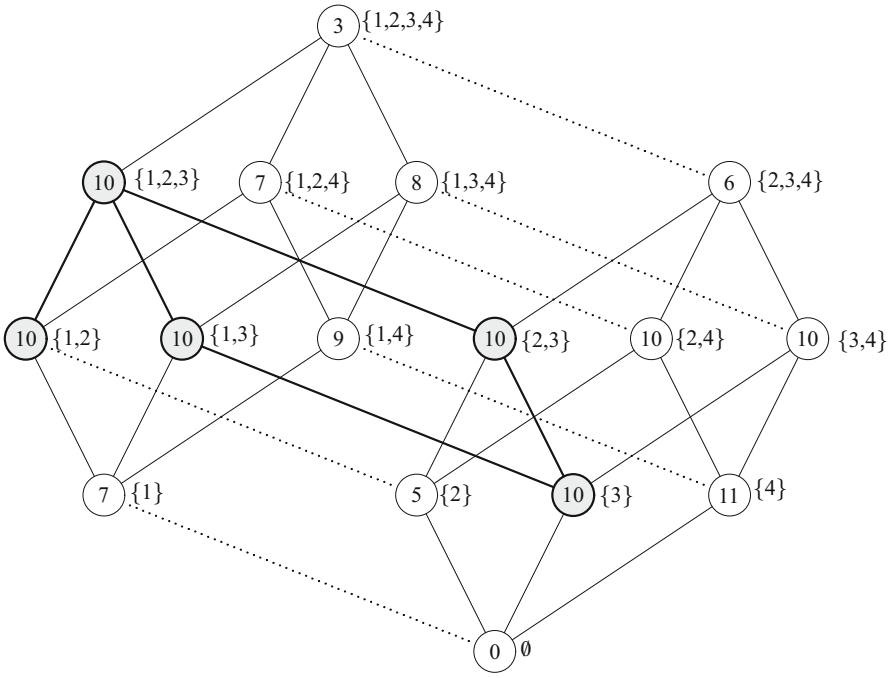

Fig. 2.6 Lower local maxima: $\{1,2\}$, $\{3\}$; upper local maximum: $\{1,2,3\}$; SDC (*shadowed*); global maximum: $\{4\}$

By the following definitions [89] (see also [64]) introduced two kinds of components of subgraphs of local maxima.

Let V_0 be the subset of V corresponding to all local maxima of z and let $H_0 = (V_0, E_0, z)$ be the subgraph of G induced by V_0. This subgraph consists of at least one component. We denote the components by $H_0^j = (V_0^j, E_0^j, z)$, with $j \in J_0 = \{1, \ldots, r\}$. Note that if L_1 and L_2 are vertices in the same component, then $z(L_1) = z(L_2)$.

A component H_0^j is called a *strict local maximum component* (STC) if for each $I \notin V_0^j$, for which there is an edge (I,L) with $L \in V_0^j$, we have $z(I) < z(L)$. A component H_0^j is called a *saddle local maximum component* (SDC) if for some $I \notin V_0^j$, there exists an edge (I,L) with $L \in V_0^j$ such that $z(I) = z(L)$. An example of the SDC defined by two intervals $[\{1,2\},\{1,2,3\}]$ and $[\{3\},\{1,2,3\}]$ is shown in Fig. 2.6. The values of a submodular function in Fig. 2.6 are printed inside the vertices. Here a trivial STC by the vertex $\{4\}$ is defined. Note that $\{3,4\}$ is not a local maximum because its neighbor $\{4\}$ is the global maximum with value $z(\{4\}) = 11$.

All vertices in a component H_0^j are local maxima of the same kind. Therefore, the index set J_0 of these components can be split into two subsets: J_1 being the index set of the STCs, and J_2 being the index set of the SDCs.

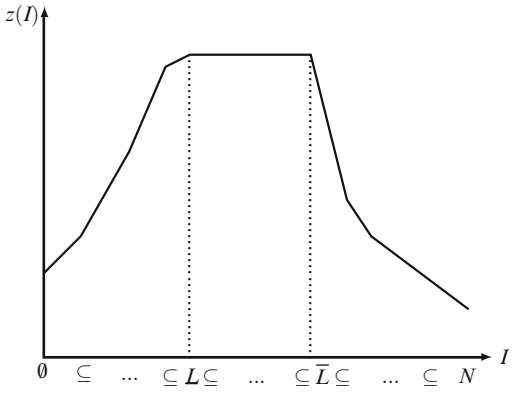

Fig. 2.7 The behavior of a submodular function on a chain with lower and upper local maxima (Khachaturov's theorem)

The following theorem of [89] is an application of Theorem 2.2 to the case of a nontrivial STC (see Fig. 2.7).

Theorem 2.3. *Let z be a submodular function on 2^N and let $\underline{L}$ and $\overline{L}$ be lower and upper maxima with $\underline{L} \subseteq \overline{L}$, both located in an STC. Then z is strictly increasing on each subchain $\emptyset \subseteq \cdots \subseteq \underline{L}$ of $[\emptyset, \underline{L}]$, constant on $[\underline{L}, \overline{L}]$, and strictly decreasing on each subchain $\overline{L} \subseteq \cdots \subseteq N$ of $[\overline{L}, N]$.*

Proof. We first show that z is strictly increasing on $[\emptyset, \underline{L}]$. The proof of the strictly decreasing case is similar. If either $\underline{L} = \emptyset$ (we obtain the decreasing case) or $|\underline{L}| = 1$, the assertion is true, since $\underline{L}$ is a local maximum of z. So, let $|\underline{L}| > 1$ and $I, J \in [\emptyset, \underline{L}]$ such that $J = I + k$, $k \in \underline{L} \setminus I$. Note that $\emptyset \subseteq I \subset J \subseteq \cdots \subseteq \underline{L}$. The submodularity of z implies $z(J) + z(\underline{L} - k) \geq z(I) + z(\underline{L})$, or $z(J) - z(I) \geq z(\underline{L}) - z(\underline{L} - k)$. Since $\underline{L} \in V_0^j$ for some $j \in J_1$, $z(\underline{L}) - z(\underline{L} - k) > 0$. Hence $z(J) > z(I)$, and we have finished the proof of the strictly increasing case.

The property that z is constant on $[\underline{L}, \overline{L}]$ follows from Corollary 2.1. □

Note that $\underline{L}$ and $\overline{L}$ need not be lower and upper maxima in Theorem 2.3. It is clear from the proof of Theorem 2.3 that any pair of embedded local maxima L_1 and L_2 located on a chain $\emptyset \subseteq \cdots \subseteq L_1 - i \subset L_1 \subseteq \cdots \subseteq L_2 \subset L_2 + k \subseteq \cdots \subseteq N$ such that $z(L_1 - i) < z(L_1)$ and $z(L_2 + k) < z(L_2)$ will imply that z is strictly increasing on each subchain $\emptyset \subseteq \cdots \subseteq L_1 - i \subset L_1$ and strictly decreasing on each subchain $L_2 \subset L_2 + k \subseteq \cdots \subseteq N$. We call such a local maximum *boundary local maximum*. In other words, a boundary local maximum is connected with vertices outside the component.

Lemma 2.2. *Let $L \in V_0^j$ for some $j \in J_1$, and let I satisfy $z(I) = z(L)$ with $(I, L) \in E$. Then $I \in V_0^j$ for the same $j \in J_1$.*

Proof. Let $L \in V_0^j$ for some $j \in J_1$. If $I \notin V_0^j$, then $z(I) < z(L)$, since $(I, L) \in E$ and L is a local maximum of the STC. □

Khachaturov [89] has observed that any global maximum belongs to a STC.

Theorem 2.4. *Let S be a global maximum of the submodular function z defined on 2^N. Then $S \in V_0^j$ for some $j \in J_1$.*

Proof. Suppose, to the contrary, that $S \in V_0^i$ with $i \in J_2$. Then there exists an $I \in V \setminus V_0$, adjacent to some $J \in V_0^i$ with $z(I) = z(J)$. This I is not a local maximum and hence I has an adjacent vertex M with $z(M) > z(I)$. Thus $z(S) = z(J) = z(I) < z(M)$, contradicting the assumption that S is a global maximum of z. □

Theorem 2.4 implies that we may restrict the search for a global maximum of a submodular function z to STCs. Based on Corollary 2.1, and definitions of strict and saddle components, we can represent each component of local maxima as a maximal connected set of intervals whose end points are lower and upper local maxima.

2.3 Excluding Rules: An Old Proof

There are two "excluding rules" (see [3, 46, 118]) that can be used to eliminate certain subsets from $[\emptyset, N]$ when determining a global maximum of a submodular function.

Theorem 2.5. *Let z be a submodular function on $[\emptyset, N]$ and V_0^j with $j \in J_0$ be the components of local maxima. Then the following assertions hold.*

(a) First Strict Excluding Rule (FSER).
If for some T_1 and T_2 with $\emptyset \subseteq T_1 \subset T_2 \subseteq N$ we have $z(T_1) > z(T_2)$, then $V_0^j \cap [T_2, T] = \emptyset$ for all $j \in J_0$.

(b) Second Strict Excluding Rule (SSER).
If for some S_1 and S_2 with $\emptyset \subseteq S_1 \subset S_2 \subseteq N$ we have $z(S_1) < z(S_2)$, then $V_0^j \cap [S, S_1] = \emptyset$ for all $j \in J_0$.

Proof. We prove case (a) because a proof of case (b) is similar. Let us consider a chain $\emptyset \subseteq \cdots \subseteq T_1 \subset T_2 \subseteq L \subseteq T \subset \ldots \subset N$ with $L \in V_0^j \cap [T_2, T] \neq \emptyset$ for some $j \in J_0$. Applying Theorem 2.2 to the subchain $\emptyset \subseteq \cdots \subseteq T_1 \subset T_2 \subseteq L$, we have $z(\emptyset) \leq \cdots \leq z(T_1) \leq z(T_2) \leq z(L)$ which contradicts $z(T_1) > z(T_2)$. □

This theorem states that by applying the strict rules we do not exclude any local maximum. In other words, we preserve all local maxima. In the following theorem of [89] we will see that applying excluding rules with nonstrict inequalities (nonstrict rules) will preserve at least one local maximum of each STC. We will call such a maximum a *representative* of the STC.

Theorem 2.6. *Let z be a submodular function on $[S, T] \subseteq [\emptyset, N]$ and for every $j \in J_1, V_0^j \cap [S, T] \neq \emptyset$. Then the following assertions hold.*

(a) First Excluding Rule (FER).
If for some T_1 and T_2 with $S \subseteq T_1 \subset T_2 \subseteq T$ holds that $z(T_1) \geq z(T_2)$, then $V_0^j \cap ([S, T] \setminus [T_2, T]) \neq \emptyset$ for all $j \in J_1$.

(b) Second Excluding Rule (SER).
If for some S_1 and S_2 with $S \subseteq S_1 \subset S_2 \subseteq T$ holds that $z(S_1) \leq z(S_2)$, then $V_0^j \cap ([S,T] \setminus [S,S_1] \neq \emptyset$ for all $j \in J_1$.

Proof. We prove case (a) because the proof of case (b) is similar. Let us consider two cases:

Case 1. $z(T_1) > z(T_2)$. Theorem 2.5 implies that $V_0^j \cap [T_2,T] = \emptyset$ for all $j \in J_0 = J_1 \cup J_2$. Since for every $j \in J_1, V_0^j \cap [S,T] \neq \emptyset$ and $[T_2,T] \subset [S,T]$ we have $([S,T] \setminus [T_2,T]) \cap V_0^j \neq \emptyset$ for all $j \in J_1$.

Case 2. $z(T_1) = z(T_2)$. If we can construct a chain through two boundary local maxima L_1 and L_2 that also contains T_1 and T_2, then we have just two possibilities:

1. $L_1 \subseteq T_1 \subset T_2 \subseteq L_2$.
2. All others.

Each case of the possibility (2) contradicts Theorem 2.3. Therefore, $L_1 \subseteq T_1 \subset T_2 \subseteq L_2$, and $L_1 \subseteq T_1 \in ([S,T] \setminus [T_2,T]) \cap V_0^j \neq \emptyset$ for all $j \in J_1$. □

In Sect. 2.6 we will give an example of the SPLP in which by application of a nonstrict excluding rule we discard the local minimum $\{2,4\}$ of the corresponding supermodular function. This local minimum is an analogue of the trivial SDC for the corresponding supermodular function.

By applying Theorem 2.6a (respectively, 2.6b) we can discard $2^{|T \setminus T_2|}$ (respectively, $2^{|S_1 \setminus S|}$) subsets of interval $[T_2,T]$ (respectively, $[S,S_1]$) because this interval does not include a local maximum of any STC from $[S,T]$. If $T_1 = S$ and $T_2 = S+i$ then in case of Theorem 2.6a the interval $[S+i,T]$ can be discarded. If $S_1 = T-i$ and $S_2 = T$ then in case of Theorem 2.6b the interval $[S,T-i]$ can be discarded. These two special cases are important because we may exclude a *half subinterval* of the current interval while we preserve at least one representative from each STC. The rules excluding a half subinterval are called *prime rules*.

Based on the last special cases of excluding rules, we present Cherenin's Preliminary Preservation (Dichotomy) Algorithm for the maximization of submodular functions in Sect. 2.5. Before we present the dichotomy algorithm we give in Theorem 2.7 an alternative proof of the correctness of these special cases of excluding rules which is based only on Lemma 2.2, the definitions of an STC and a submodular function z. This proof shows that in case of submodular functions the definition of a STC is an insightful notion for understanding the correctness of Cherenin's dichotomy algorithm. Therefore, it is not necessary to use all the statements of the previous section in order to justify both prime rules. In the next section we present a generalization and a simple justification of the same rules.

Theorem 2.7. *Let z be a submodular function on 2^N. Suppose that for $\emptyset \subseteq S \subset T \subseteq N$ and for every $j \in J_1$, $V_0^j \cap [S,T] \neq \emptyset$. Then the following assertions hold.*

(a) First Prime Excluding Rule (FPER).
If for some $i \in T \setminus S$ it holds that $z(S+i) \leq z(S)$, then $[S,T-i] \cap V_0^j \neq \emptyset$ for all $j \in J_1$.

(b) Second Prime Excluding Rule (SPER).
If for some $i \in T \setminus S$ it holds that $z(T-i) \leq z(T)$, then $[S+i,T] \cap V_0^j \neq \emptyset$ for all $j \in J_1$.

Proof. We prove only part (a). The proof of part (b) is similar.

(a) Let $z(S+i) \leq z(S)$ for some $i \in T \setminus S$ and let $G \in V_0^j \cap [S,T]$ for any $j \in J_1$. Then $S \subset G$.

Case 1. $i \in G$. From the definition of submodularity applied to the sets $G-i$ and $S+i$
$z(G-i)+z(S+i) \geq z(G \cup S+i)+z(S) \Rightarrow$
$z(G-i)-z(G \cup S+i) \geq z(S)-z(S+i) \geq 0 \Rightarrow$
$z(G-i) \geq z(G \cup S+i) = z(G) \Rightarrow$ (G is a local maximum)
$z(G-i) = z(G)$. $G \in V_0^j \Rightarrow$ (by Lemma 2.2)
$G-i \in V_0^j \Rightarrow G-i \in V_0^j \cap [S,T-i] \Rightarrow V_0^j \cap [S,T-i] \neq \emptyset$.

Case 2. $i \notin G$.
$i \notin G \Rightarrow G \in V_0^j \cap [S,T-i] \Rightarrow V_0^j \cap [S,T-i] \neq \emptyset$. □

Theorem 2.7a states that if $z(S+i)-z(S) \leq 0$ for some $i \in T \setminus S$, then by preserving the interval $[S,T-i]$ we preserve at least one strict local maximum from each STC, and hence we preserve at least one global maximum from each STC containing a global maximum. Therefore, in this case it is possible to exclude exactly the whole interval $[S+i,T]$ of $[S,T]$ from consideration when searching for a global maximum of the submodular function z on $[S,T] \subseteq [\emptyset,N]$. For example, see Fig. 2.8, if $z(\emptyset)-z(\emptyset+1) \geq 0$, then the interval $[\{1\},\{1,2,3,4\}]$ can be excluded, i.e., the interval $[\emptyset,\{2,3,4\}]$ will be preserved (FPER). If $z(\{1,2,3,4\})-z(\{1,2,3,4\}-1) \geq 0$, then the interval $[\emptyset,\{2,3,4\}]$ can be excluded, i.e., the interval $[\{1\},\{1,2,3,4\}]$ will be preserved (SPER).

If the prime rules are not applicable it will be useful to discard less than a half subinterval of the current interval $[S,T] \subseteq [\emptyset,N]$. In the following section we further relax most of the theoretical results presented in the previous sections of this chapter with the purpose to show the correctness of all excluding rules and their generalizations (preservation rules) based only on the definitions of submodularity and the maximum value $z^*[S,T]$ of the function z on the interval $[S,T] \subseteq [\emptyset,N]$.

2.4 Preservation Rules: Generalization and a Simple Justification

In the following theorem we give an important interpretation of the submodularity property which is based on two pairs of submodular function values. For this purpose we introduce an *upper (respectively, lower) partition* of the current interval $[S,T]$ by an inner vertex $Q : S \subset Q \subset T$ into two parts $[Q,T]$ and $[S,T] \setminus [Q,T]$

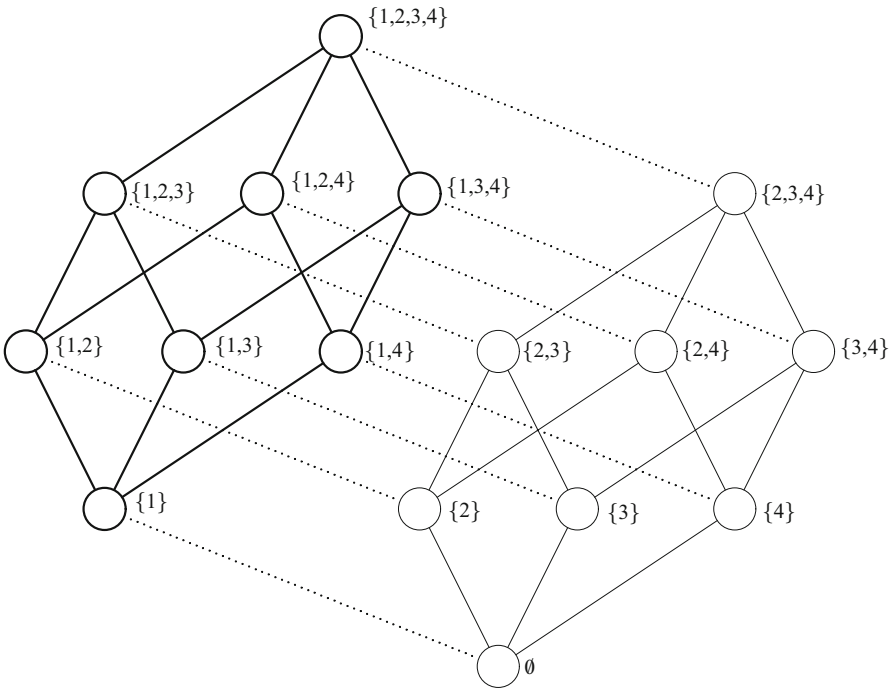

Fig. 2.8 Example of prime excluding rules

(respectively, $[S,Q]$ and $[S,T] \setminus [S,Q]$). In terms of the maximum values of the function z defined on each of two parts of the above-mentioned partitions, a special case of submodularity can be read as either $z^*([S,T] \setminus [Q,T]) + z(Q) \geq z(S) + z^*[Q,T]$ or $z^*([S,T] \setminus [S,Q]) + z(Q) \geq z^*[S,Q] + z(T)$.

Here, the both maximal values of a submodular function and their arguments (vertices) involved in each of the above indicated inequalities are *unknown*. In other words, Theorem 2.8 establishes a relationship of the difference between the unknown optimal values of z on the two parts of the partition, for example, $([S,T] \setminus [Q,T])$ and $[Q,T]$ of $[S,T]$ and the corresponding difference $z(S) - z(Q)$ (see the FER in Theorem 2.6); a symmetrical result is obtained for the SER.

Theorem 2.8. *Let z be a submodular function on the interval $[S,T] \subseteq [\emptyset, N]$. Then for any Q such that $S \subset Q \subset T$ the following assertions hold.*

(a) $z^*([S,T] \setminus [Q,T]) - z^*[Q,T] \geq z(S) - z(Q)$.
(b) $z^*([S,T] \setminus [S,Q]) - z^*[S,Q] \geq z(T) - z(Q)$.

Proof. We prove only case (a) because the proof of case (b) is similar. Let $z^*[Q,T] = z(Q \cup J)$ with $J \subseteq T \setminus Q$. Define $I = S \cup J$. Then $I \in [S,T] \setminus [Q,T]$ since $Q \setminus S \not\subseteq I$. We have that $z^*([S,T] \setminus [Q,T]) - z(S) \geq z(I) - z(S) = z(S \cup J) - z(S)$. From the submodularity of z we have $z(S \cup J) - z(S) \geq z(Q \cup J) - z(Q)$. Therefore, $z^*([S,T] \setminus [Q,T]) - z(S) \geq z^*[Q,T] - z(Q)$. □

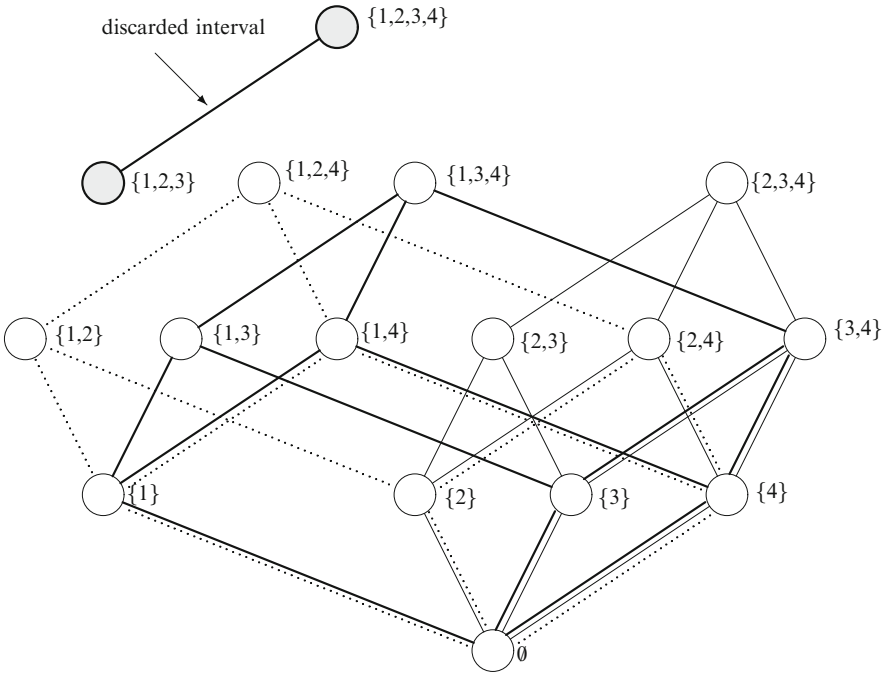

Fig. 2.9 A representation of the upper partition of the interval $[S,T] = [\emptyset, \{1,2,3,4\}]$ with $Q \setminus S = \{1,2,3\}$.

Theorem 2.8 is a generalization of Cherenin–Khachaturov's rules stating that the difference of values of a submodular function on any pair of embedded subsets is a lower bound for the difference between the optimal values of z on the two parts of the partition defined by this pair of embedded subsets. The theorem can be used to decide in which part of the partition $([S,T] \setminus [Q,T])$ and $[Q,T]$ of $[S,T]$ a global maximum of z is located.

We may represent the partition of interval $[S,T]$ from Theorem 2.8 by means of its proper subintervals as follows:

$$(a)\ \textit{upper partition}\ [S,T] \setminus [Q,T] = \bigcup_{i \in Q \setminus S} [S, T-i]$$

and

$$(b)\ \textit{lower partition}\ [S,T] \setminus [S,Q] = \bigcup_{i \in T \setminus Q} [S+i, T].$$

Examples of upper and lower partitions in Figs. 2.9 and 2.10 are shown.

A disadvantage of representations (a) and (b) is a nonempty overlapping of each pairwise distinct intervals involved in these representations. As easy to see in Figs. 2.11 and 2.12 we can avoid such an overlapping by representing the remaining

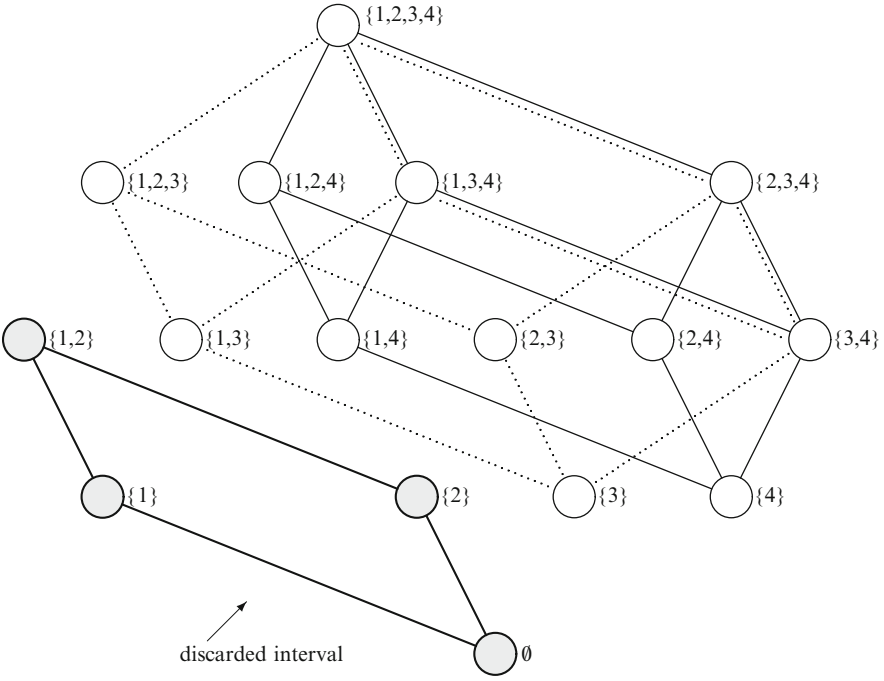

Fig. 2.10 A representation of the lower partition by $Q = \{1,2\}$ for the interval $[S,T] = [\emptyset,\{1,2,3,4\}]$ with $T \setminus Q = \{3,4\}$

parts $([S,T] \setminus [Q,T])$ and $([S,T] \setminus [S,Q])$ with a sequence of "parallel" nonoverlapping intervals. For example, the difference $[\emptyset,\{1,2,3,4\}] \setminus [\{1,2,3\},\{1,2,3,4\}] = [\{1,2\},\{1,2,4\}] \cup [\{1\},\{1,3,4\}] \cup [\emptyset,\{2,3,4\}]$ (see Figs. 2.9 and 2.10), and the difference $[\emptyset,\{1,2,3,4\}] \setminus [\emptyset,\{1,2\}] = [\{3\},\{1,2,3\}] \cup [\{4\},\{1,2,3,4\}]$ (see Figs. 2.11 and 2.12).

The sequence of nonoverlapping intervals can be created by the following iterative procedure. We will use the value $d = \dim([U,W])$ of the *dimension* of an interval $[U,W]$ interpreted as the corresponding subspace of the Boolean space $\{0,1\}^n$ which is another representation of the interval $[\emptyset,N]$.

If we have discard the k-dimensional subinterval $[Q,T]$ in the upper partition of the interval $[S,T]$, then the first nonoverlapping interval $[U_1,W_1]$ is the k-dimensional subinterval of the $(k+1)$-dimensional interval $[U_1,T] = [Q,T] \cup [U_1,W_1]$. In other words, the first nonoverlapping interval $[U_1,W_1]$ is the k-dimensional complement to the $(k+1)$-dimensional interval $[U_1,T]$ such that $[U_1,W_1] = [U_1,T] \setminus [Q,T]$. The second nonoverlapping interval $[U_2,W_2]$ is the $(k+1)$-dimensional subinterval of the $(k+2)$-dimensional interval $[U_2,T] = [U_1,T] \cup [U_2,W_2]$, and $[U_2,W_2] = [U_2,T] \setminus [U_1,T]$, etc. Finally, $[U_q,W_q] = [U_q,T] \setminus [U_{(q-1)},T]$. The number q of the nonoverlapping intervals in the upper partition is equal to $n-k$, where $k = \dim[Q,T]$. The representation of a lower partition by the sequence of nonoverlapping intervals

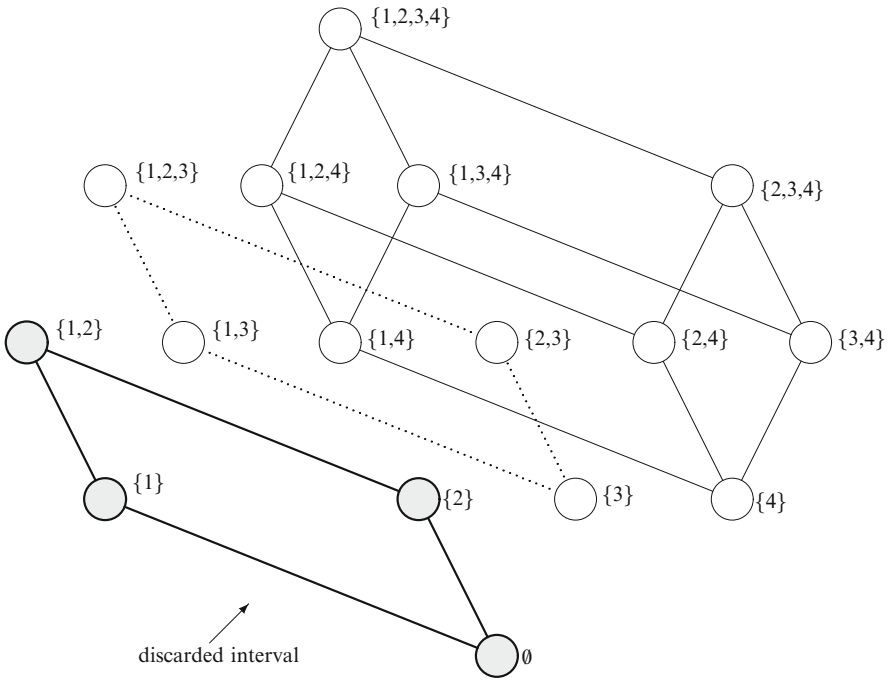

Fig. 2.11 The non-overlapping representation of the lower partition by parallel intervals $[\{3\},\{1,2,3\}]$ and $[\{4\},\{1,2,3,4\}]$.

can be described in similar lines. Note that the above indicated representation of lower (upper) partition by a sequence of nonoverlapping intervals has the *minimum number of mutually disjoint intervals*.

For example (see Fig. 2.12), the complement interval to $[\{1,2,3\},\{1,2,3,4\}]$ is $[\{1,2\},\{1,2,4\}]$ since $[\{1,2\},\{1,2,4\}]\cup[\{1,2,3\},\{1,2,3,4\}]=[\{1,2\},\{1,2,3,4\}]$, and the complement to $[\{1,2\},\{1,2,3,4\}]$ is $[\{1\},\{1,3,4\}]$. Finally, the complement to $[\{1\},\{1,2,3,4\}]$ is $[\emptyset,\{2,3,4\}]$.

If, in Theorem 2.8, we replace Q by $S+k$ in part (a), and Q by $T-k$ in part (b), we obtain the following generalization of the prime rules stated in Theorem 2.7.

Corollary 2.2. *Let z be a submodular function on the interval $[S,T]\subseteq[\emptyset,N]$ and let $k\in T\setminus S$. Then the following assertions hold.*

(a) $z^*[S,T-k]-z^*[S+k,T]\geq z(S)-z(S+k)$.
(b) $z^*[S+k,T]-z^*[S,T-k]\geq z(T)-z(T-k)$.

By adding the condition $z(S)-z(S+k)\geq 0$ to part (a) and the condition $z(T)-z(T-k)\geq 0$ to part (b) of Corollary 2.2, we obtain another form (see Corollary 2.3) of two prime rules from Theorem 2.7 for preserving subintervals containing at least one global maximum of z on $[S,T]$.

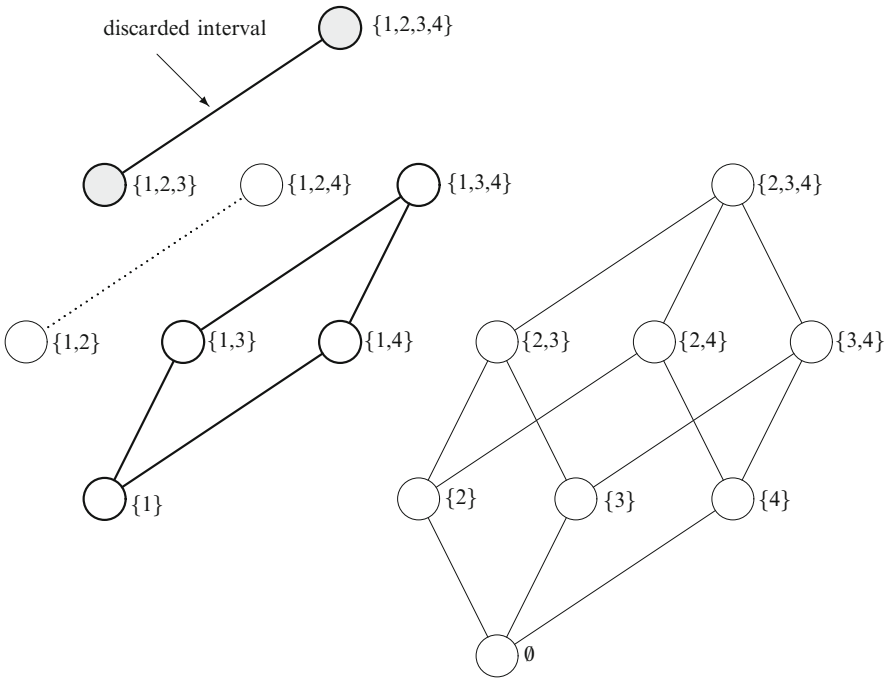

Fig. 2.12 The nonoverlapping representation of the upper partition by the parallel intervals $[\{1,2\},\{1,2,4\}]$, $[\{1\},\{1,3,4\}]$, and $[\emptyset,\{2,3,4\}]$

Corollary 2.3. *Let z be a submodular function on the interval $[S,T] \subseteq [\emptyset,N]$ and $k \in T \setminus S$. Then the following assertions hold.*

(a) First Preservation (FP) Rule.
If $z(S) \geq z(S+k)$, then $z^*[S,T] = z^*[S,T-k] \geq z^*[S+k,T]$.
(b) Second Preservation (SP) Rule.
If $z(T) \geq z(T-k)$, then $z^*[S,T] = z^*[S+k,T] \geq z^*[S,T-k]$.

Proof. (a) From Corollary 2.2a we have $z^*[S,T-k] - z^*[S+k,T] \geq z(S) - z(S+k)$. By assumption $z(S) - z(S+k) \geq 0$. Hence, $z^*[S,T] = z^*[S,T-k] \geq z^*[S+k,T]$.
(b) The proof is similar. □

From the calculation point of view these rules are the same as in Theorem 2.6, but Theorem 2.7 is more powerful than Corollary 2.3. In Theorem 2.7 we preserve at least one strict local maximum from each STC, and hence one global maximum from each STC that contains global maxima. Corollary 2.3 only states that we preserve at least one global maximum. However, we can use Corollary 2.3 for constructing some extension of the preservation rules.

For $\varepsilon \geq 0$, we may consider the problem of finding an approximate solution $J \in [S,T]$ such that $z^*[S,T] \leq z(J) + \varepsilon$; J is called an *ε-maximum* of z on $[S,T]$.

The following corollary presents an extension of the rules from Corollary 2.3 which is appropriate to the problem of ε-maximization.

Corollary 2.4. *Let z be a submodular function on the interval $[S,T] \subseteq [\emptyset,N]$, and $k \in T \setminus S$. Then the following assertions hold.*

(a) First θ-Preservation (θ-FP) Rule.
If $z(S)-z(S+k)=\theta<0$, then $z^*[S,T]-z^*[S,T-k] \leq -\theta$, which means that $[S,T-k]$ contains a $|\theta|$-maximum of $[S,T]$.

(b) Second η-Preservation (η-SP) Rule.
If $z(T)-z(T-k)=\eta<0$, then $z^*[S,T]-z^*[S+k,T] \leq -\eta$, which means that $[S+k,T]$ contains a $|\eta|$-maximum of $[S,T]$.

Proof. The proof of part (a) is as follows.

Case 1. If $z^*[S,T]=z^*[S,T-k]$ then $z^*[S,T-k]-z^*[S,T-k] \leq -\theta$ or $z^*[S,T]-z^*[S,T-k] \leq -\theta$.

Case 2. If $z^*[S,T]=z^*[S+k,T]$, then from Theorem 2.7a follows that $z^*[S,T-k]-z^*[S+k,T] \geq \theta$ or $z^*[S,T-k]-z^*[S,T] \geq \theta$. Hence $z^*[S,T]-z^*[S,T-k] \leq -\theta$. The proof of (b) is similar. □

2.5 The PPA

By means of Corollary 2.3 it is often possible to exclude a large part of $[\emptyset,N]$ from consideration when determining a global maximum of z on $[\emptyset,N]$. The so-called *PPA* (see [66]) determines the smallest subinterval $[S,T]$ of $[\emptyset,N]$ containing a global maximum of z, by using the preservation rules of Corollary 2.3.

We call the PPA the *dichotomy algorithm* because in every successful step it halves the current domain of a submodular function.

Let $[S,T]$ be an interval. For each $i \in T \setminus S$, define $\delta^+(S,T,i)=z(T)-z(T-i)$ and $\delta^-(S,T,i)=z(S)-z(S+i)$; moreover, define $\delta^+_{\max}(S,T)=\max\{\delta^+(S,T,i) \mid i \in T \setminus S\}$, $r^+(S,T)=\min\{r \mid \delta^+(S,T,r)=\delta^+_{\max}(S,T)\}$. Similarly, for $\delta^-(S,T,i))$ define $\delta^-_{\max}(S,T)=\max\{\delta^-(S,T,i)) \mid i \in T \setminus S\}$, $r^-(S,T)=\min\{r \mid \delta^-(S,T,r)=\delta^-_{\max}(S,T)\}$. If no confusion is likely, we briefly write r^-, r^+, δ^-, δ^+ instead of $r^-(S,T)$, $r^+(S,T)$, $\delta^-_{\max}(S,T)$, and $\delta^+_{\max}(S,T)$ respectively (Fig. 2.13).

Each time that either S or T are updated during the execution of the PPA, the conditions of Corollary 2.3 remain satisfied, and therefore $z^*[S,T]=z^*[U,W]$ with $[U,W] \subseteq [S,T]$ remains invariant at each step of the PPA. At the end of the algorithm we have that $\max\{\delta^+,\delta^-\}<0$, and therefore $z(S)<z(S+i)$ and $z(T)<z(T-i)$ for each $i \in T \setminus S$. Hence Corollary 2.3 cannot be applied to further reduce the interval $[S,T]$ without violating $z^*[S,T]=z^*[U,W]$. Note that this remark shows the correctness of the procedure PP(.).

If we replace in the PPA the rules of Corollary 2.3 by those of Corollary 2.4, we obtain an ε-maximization variant of the PPA. In this case the output of the

Procedure PP(U,W,S,T)
Input: A submodular function z on the subinterval $[U,W]$ of $[0,N]$
Output: A subinterval $[S,T]$ of $[U,W]$ such that $z^*[S,T]=z^*[U,W]$, $z(S)<z(S+i)$ and $z(T)<z(T-i)$ for each $i\in T\setminus S$.

```
begin
    S ← U;    T ← W;
    Step 1: if S = T
        then goto Step 4;
    Step 2: Calculate δ+ and r+;
        if δ+ ≥ 0 (Corollary 2.3b)
        then begin call PP(S + r+, T; S, T)
                goto Step 4
            end;
    Step 3: Calculate δ− and r−;
        if δ− ≥ 0 (Corollary 2.3a)
        then begin call PP(S, T − r−; S, T)
                goto Step 4
            end;
    Step 4:
end;
```

Fig. 2.13 The dichotomy (preliminary preservation) algorithm

ε-PPA will be a subinterval $[S,T]$ of $[U,W]$ such that $z^*[U,W]-z^*[S,T]\le\varepsilon$ with postconditions $z(S)+\varepsilon<z(S+i)$ and $z(T)+\varepsilon<z(T-i)$ for each $i\in T\setminus S$.

The following theorem can also be found in [58, 66]. It provides an upper bound for the worst case complexity of the PPA; the complexity function is dependent only on the number of comparisons of pairs of values for z.

Theorem 2.9. *The time complexity of the PP algorithm procedure is at most* $O(n^2)$.

Note that if the PPA terminates with $S=T$, then S is a global maximum of z. Any submodular function z on $[U,W]$ for which the PP algorithm returns a global maximum for z is called a *PP-function*.

An example of a set of PP-functions $\mathscr{P}$ is shown in Fig. 2.14. Here, for all vertices without prespecified values of $z(I)$ can be assigned an arbitrary value of $z(I)$ such that each corresponding set function $z(I)\in\mathscr{P}$ defined on the whole weighted graph G will be submodular. For example, if for all vertices without prespecified values of $z(I)$ in Fig. 2.14 we set $z(I)=a$, then for each real valued constant $a:2\le a\le 3$ the corresponding function z is a submodular PP-function. It means that by applying the dichotomy algorithm we have found an optimal solution to the PSMF for all PP-functions defined by a constant a.

Corollary 2.5 describes in terms of STCs some properties of the variables S and T during the iterations of the PPA. A representative $L_1^j\in V_0^j$ with $j\in J_1$ which will be preserved through all iterations during the execution of the PPA by FPER

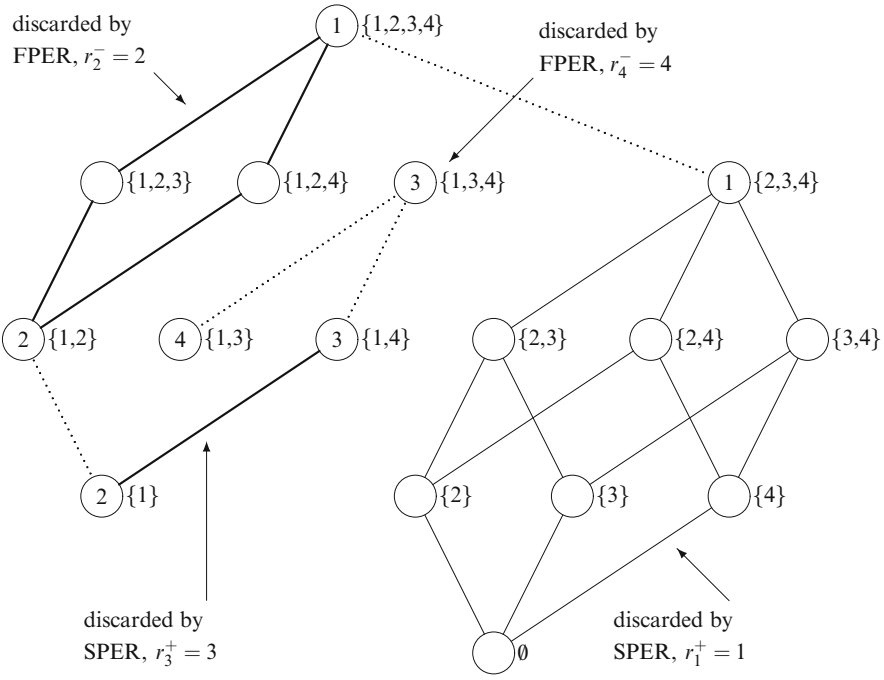

Fig. 2.14 The idea of the dichotomy algorithm: $z(\{1,3\}) = 4$ is the global maximum for all submodular functions from the subclass of $\mathscr{P}$

($L_1^j \in V_0^j \cap [S, T-i] \neq \emptyset$ with $j \in J_1$) or SPER ($L_1^j \in V_0^j \cap [S+i, T] \neq \emptyset$ with $j \in J_1$) is called a *PP-representative* of STC H_0^j with $j \in J_1$ (see Theorem 2.7).

Corollary 2.5. *If z is a submodular PP-function on $[U,W] \subseteq [\emptyset, N]$, then at each iteration of the PPA $S \subseteq \cap_{j\in J_1} L_1^j$ and $T \supseteq \cup_{j\in J_1} L_1^j$.*

Proof. Theorem 2.7a says that if $z(S+i) - z(S) \leq 0$ for some $i \in T \setminus S$, then by preserving the interval $[S, T-i]$ we preserve at least one PP-representative L_1^j from each STC H_0^j, and hence $i \notin L_1^j$. In case of Theorem 2.7b we preserve PP-representatives L_1^j such that $i \in L_1^j$ for all STCs in $[S,T]$. Therefore, $i \in S \subseteq \cap_{j\in J_1} L_1^j$ and $T \supseteq \cup_{j\in J_1} L_1^j$. □

The following theorem gives a property of PP-functions in terms of STCs.

Theorem 2.10. *If z is a submodular PP-function on $[U,W] \subseteq [\emptyset, N]$, then $[U,W]$ contains exactly one STC.*

Proof. From $\cap_{j\in J_1} L_1^j \supseteq S = T \supseteq \cup_{j\in J_1} L_1^j$ we obtain $\cap_{j\in J_1} L_1^j = \cup_{j\in J_1} L_1^j$ or $L_1^j = L$ for all $j \in J_1$. □

Note that not each submodular function with exactly one STC on $[\emptyset, N]$ is a PP-function. For example, let $N = \{1,2,3\}$ and consider the submodular function z defined by $z(I) = 2$ for any $I \in [\emptyset, \{1,2,3\}] \setminus (\{\emptyset\} \cup \{1,2,3\})$ and $z(I) = 1$ for $I \in (\{\emptyset\} \cup \{1,2,3\})$. The vertex set of the unique STC defined by this function can be represented by $[\{1\},\{1,2\}] \cup [\{1\},\{1,3\}] \cup [\{2\},\{1,2\}] \cup [\{2\},\{2,3\}] \cup [\{3\},\{1,3\}] \cup [\{3\},\{2,3\}]$. The PPA terminates with $[S,T] = [\emptyset,\{1,2,3\}]$ and so, z is not a PP-function.

2.6 Nonbinary Branching Rules

Usually in BnB type algorithms we use a *binary* branching rule by which the original set $[S,T]$ of feasible solutions will be split by an element k into two subsets $[S+k,T]$ and $[S,T-k]$. Let us consider an interval $[S,T]$ for which the postconditions of the PPA are satisfied, i.e., $z(S) < z(S+i)$ and $z(T) < z(T-i)$ for each $i \in T \setminus S$. Thus the PPA cannot make the interval $[S,T]$ smaller. By using Corollary 2.6 we can sometimes find two subintervals $[S,T-k_1]$ and $[S,T-k_2]$ such that the postconditions of the PPA algorithm for each of these intervals are violated.

Corollary 2.6. *Let z be a submodular function on the interval $[S,T] \subseteq [\emptyset,N]$ and let $k_1,k_2 \in T \setminus S$ with $k_1 \neq k_2$. Then the following assertions hold.*

(a) $\max\{z^*[S,T-k_1], z^*[S,T-k_2]\} - z^*[S+k_1+k_2,T] \geq z(S) - z(S+k_1+k_2)$.
(b) $\max\{z^*[S+k_1,T], z^*[S+k_2,T]\} - z^*[S,T \setminus \{k_1,k_2\}] \geq z(T) - z(T \setminus \{k_1,k_2\})$.

Proof. We prove only part (a) because the proof of part (b) is similar. Replace Q by $S+k_1+k_2$ in Theorem 2.8a. Then, $z^*([S,T] \setminus [Q,T]) - z^*[Q,T] = z^*(\bigcup_{i \in Q \setminus S}[S,T-i]) - z^*[Q,T] = z^*([S,T-k_1] \cup [S,T-k_2]) - z^*[S+k_1+k_2,T] = \max\{z^*[S,T-k_1], z^*[S,T-k_2]\} - z^*[S+k_1+k_2,T] \geq z(S) - z(Q) = z(S) - z(S+k_1+k_2)$. □

In the case that $z(S) - z(S+k_1+k_2) \geq 0$ we can discard the interval $[S+k_1+k_2,T]$ and continue the search for an optimal solution by applying the PPA separately to each remaining interval $[S,T-k_1]$ and $[S,T-k_2]$, which are obtained by subtracting an element k_i from T. The symmetrical case will be obtained if $z(T) - z(T \setminus \{k_1,k_2\}) \geq 0$. Corollary 2.6 can easily be generalized to the case of m-ary branching by elements $k_1,k_2,\ldots,k_m$ with $m \leq |T \setminus S|$.

We conclude this section with a simple plant location example borrowed from [18]. The data are presented in Table 2.1.

For solving the SPLP it suffices to solve the problem $\min\{z(I) \mid I \in [\emptyset,N]\} = z^*[\emptyset,N] = z(G)$ with $N = \{1,2,3,4\}$, $m = 5$ and

$$z(I) = \sum_{i \in I} f_i + \sum_{j=1}^{m} \min_{i \in I} c_{ij}.$$

Table 2.1 The data of the SPLP

Location		Delivery cost to site				
i	f_i	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
1	7	7	15	10	7	10
2	3	10	17	4	11	22
3	3	16	7	6	18	14
4	6	11	7	6	12	8

As usual for the SPLP, f_i is the fixed cost of opening a plant at location i, c_{ij} is the cost of satisfying the demand of customer j by plant i, and $z(I)$ is a supermodular function. Note that if in the definition of a submodular function we change the sign "$\geq$" to the opposite sign "$\leq$," then we obtain the definition of a supermodular function. For the sake of completeness, let us show that $z(I)$ of the SPLP is supermodular

Lemma 2.3. *The objective $z(I)$ of the SPLP is supermodular.*

Proof. According to Theorem 2.1(i) a function is supermodular if

$$z(A)+z(B)\leq z(A\cup B)+z(A\cap B),\ \ \forall A,B\subseteq N.$$

We use the following representation of this definition

$$z(A)+z(B)-z(A\cup B)+z(A\cap B)\leq 0,\ \ \forall A,B\subseteq N.$$

Substituting

$$z(I)=\sum_{i\in I} f_i+\sum_{j=1}^{m}\min_{i\in I} c_{ij}$$

gives

$$\begin{aligned}&\textstyle\sum_{i\in A} f_i+\sum_{j=1}^{m}\min_{i\in A} c_{ij}+\sum_{i\in B} f_i+\sum_{j=1}^{m}\min_{i\in B} c_{ij}\\ &\quad-\textstyle\sum_{i\in A\cup B} f_i-\sum_{j=1}^{m}\min_{i\in A\cup B} c_{ij}-\sum_{i\in A\cap B} f_i-\sum_{j=1}^{m}\min_{i\in A\cap B} c_{ij}\\ &\quad=\textstyle\sum_{i\in A} f_i+\sum_{i\in B} f_i-\sum_{i\in A\cup B} f_i-\sum_{i\in A\cap B} f_i\\ &\quad+\textstyle\sum_{j=1}^{m}[(\min_{i\in A} c_{ij}-\min_{i\in A\cup B} c_{ij})+(\min_{i\in B} c_{ij}-\min_{i\in A\cap B} c_{ij})].\end{aligned}$$

Note that

$$\left[\sum_{i\in A} f_i+\sum_{i\in B} f_i-\sum_{i\in A\cup B} f_i-\sum_{i\in A\cap B} f_i\right]=0,$$

hence it is enough to show that for each $j=1,\ldots,m$

$$[(\min_{i\in A} c_{ij}-\min_{i\in A\cup B} c_{ij})+(\min_{i\in B} c_{ij}-\min_{i\in A\cap B} c_{ij})]\leq 0.$$

Let us consider two cases.

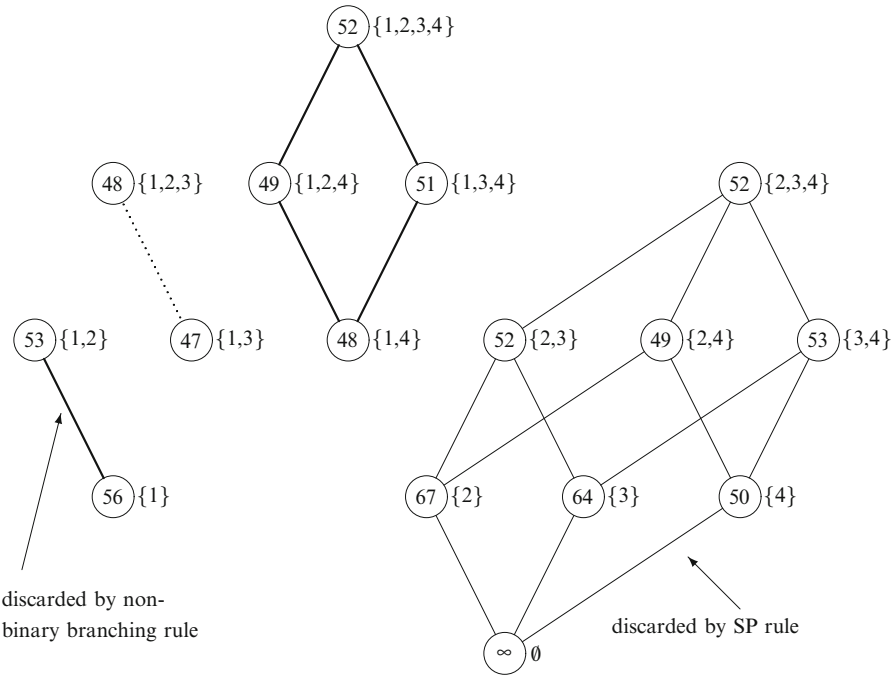

Fig. 2.15 The SPLP example: illustration of nonbinary branching rule

Case 1. $\min_{i\in A\cup B} c_{ij} = c_{aj}$ for some $a \in A$. Then $\min_{i\in A} c_{ij} = \min_{i\in A\cup B} c_{ij}$ and $\min_{i\in B} c_{ij} \leq \min_{i\in A\cap B} c_{ij}$.

Case 2. $\min_{i\in A\cup B} c_{ij} = c_{bj}$ for some $b \in B$. Then $\min_{i\in B} c_{ij} = \min_{i\in A\cup B} c_{ij}$ and $\min_{i\in A} c_{ij} \leq \min_{i\in A\cap B} c_{ij}$. □

We use this example for illustrating that the supermodular function defined by data from Table 2.1 is not a PP-function. Of course, here we mean the corresponding definition of a PP-function obtained by replacing the definitions of local, global maxima of a submodular function by the local, global minima of a supermodular function. It is easy to check that this supermodular function has two trivial analogues of STCs: $\{1,4\}$, $\{1,3\}$ and one trivial analogue of SDC: $\{2,4\}$ (see Fig. 2.15).

After the first execution of Step 3 of the PPA, we have that $[S,T] = [\{1\},\{1,2,3,4\}]$, because $\delta^+ = z(\{1,2,3,4\}) - z(\{2,3,4\}) = 0$ and $r^+ = 1$. Together with interval $[\{\emptyset\},\{2,3,4\}]$ the PPA has discarded the trivial SDC $\{2,4\}$. After the second execution of Steps 2 and 3 the PPA terminates with interval $[S,T] = [\{1\},\{1,2,3,4\}]$, because all postconditions of the PPA are satisfied. Hence, this function is not a PP-function. A global minimum of this SPLP can be found by application of the following analogue of the inequality from Corollary 2.6b:

$$\begin{aligned}&\min\{z^*[S+k_1,T], z^*[S+k_2,T]\} - z^*[S,T\setminus\{k_1,k_2\}]\\ &\quad\leq z(T) - z(T\setminus\{k_1,k_2\}).\end{aligned}$$

Let us substitute all possible pairs $\{k_1, k_2\}$ into the right-hand side of this inequality with $S = \{1\}$ and $T = \{1,2,3,4\}$. Then, we have that only $z(\{1,2,3,4\}) - z(\{1,2,3,4\} - \{3,4\}) = 52\text{–}53 < 0$. Hence, we can discard the interval $[\{1\}, \{1,2,3,4\} - \{3,4\}]$ and we may continue to find $z^*[\{1\}, \{1,2,3,4\}]$ by solving two remaining subproblems $z^*[\{1,3\}, \{1,2,3\}]$ and $z^*[\{1,4\}, \{1,2,3,4\}]$ defined on "parallel" intervals $[\{1,3\}, \{1,2,3\}]$ and $[\{1,4\}, \{1,2,3,4\}]$ (with disjoint set of feasible solutions) instead of two corresponding subproblems $z^*[S + k_1, T] = z^*[\{1,3\}, \{1,2,3,4\}]$ and $z^*[S + k_2, T] = z^*[\{1,4\}, \{1,2,3,4\}]$ which have the nonempty intersection on $[\{1,3,4\}, \{1,2,3,4\}]$. Each of these subproblems can be solved by the corresponding analogue of the PPA.

2.7 Concluding Remarks

We have considered a submodular function z defined on the Boolean hypercube to which we can apply a classic theorem of Cherenin that z is quasiconcave on any chain that intersects a local maxima component. This result enables a clearer understanding of the structure of a submodular function in terms of components of the graph of local maxima. Specifically we may state that each component of the graph of local maxima is a maximal connected set of intervals whose end points are lower and upper local maxima. Cherenin's theorem provides a justification of "the method of successive calculations." This method was successfully applied to solve problems arising in railway logistics planning (see [27, 28, 118]), and for constructing BnB type algorithms (see [3,46,56,57,60–64,66,72,89,90]) for solving a number of NP-hard problems.

We have shown that if the dichotomy algorithm (PPA) terminates with $S = T$, then the given submodular function has exactly one strict component of local maxima (STC). Hence the number of subproblems created in a branch without bounds type algorithm, which is based on the dichotomy algorithm, can be used as an upper bound for the number of the STCs. In a similar way, an upper bound for the number of all components (STCs and SDCs) by using strict excluding rules can be calculated. This information can be used for complexity analysis in terms of the number of local optima for a specific class of problems arisen in practice (computational experiments).

We next proposed a generalization of Cherenin's excluding rules given by Theorem 2.8 which provides implicit enumeration bounds for a recursive implementation of any BnB procedure incorporating the dichotomy algorithm. This generalization is useful in two respects. First, it is suitable for use in ε-*optimal* procedures which obtain an approximate global maximum within specified bounds. Second, the theorem allows the derivation of alternatives to the prime excluding rules by which we are able to discard subintervals of smaller cardinality than half original subinterval. We show that the remaining part of the current interval can be represented by a set of subintervals, some of which may include just one strict

component. In other words, we try to prepare the necessary conditions for the dichotomy algorithm to terminate on each subinterval. Moreover, Theorem 2.8 is based only on the definition of the maximum value of PMSF for an interval of $[\emptyset, N]$, and relaxed Cherenin–Khachaturov's theory presented in Sects. 2.2 and 2.3 (which is based on notions of monotonicities on a chain, local and global maxima, strict and saddle components in the Hasse diagram).

Corollary 2.2 can be considered as the basis of our data correcting (DC) algorithm presented in the next chapter. It states that if an interval $[S,T]$ is split into $[S,T-k]$ and $[S+k,T]$, then the difference between the submodular function values $z(S)$ and $z(S+k)$, or between the values of $z(T)$ and $z(T-k)$ is an upper bound for the difference of the (unknown!) optimal values on the two subintervals. This difference is used for "correcting" the current data (values of a submodular function z) in the DC algorithm. In the next chapter our computational experiments with the QCP problem show that we can substantially reduce the calculation time for data correcting algorithms [59,66] by recursive application of our main theorem.

An interesting subject for future research is the investigation of the computational efficiency of m-ary branching rules (see Corollary 2.6) for specific problems which can be reduced to the maximization of submodular functions.

Another purpose of this chapter is to present our main result [73] which can be stated as follows. For any partition of the current Hasse subdiagram $[S,T]$ spanned on a pair of embedded subsets $\emptyset \subseteq S \subset Q \subset T \subseteq N = \{1,2,\ldots,n\}$ into two parts either $[S,T] \setminus [Q,T]$ and $[Q,T]$ or $[S,T] \setminus [S,Q]$ and $[S,Q]$ defined by an inner subset Q from this subdiagram, the difference of two corresponding function values either $z(S) - z(Q)$ or $z(T) - z(Q)$ is a lower bound for the difference between the unknown(!) optimal values either $z^*([S,T] \setminus [Q,T]) - z^*([Q,T])$ or $z^*([S,T] \setminus [S,Q]) - z^*([S,Q])$, respectively, of the submodular function z. The main result was successfully used as a base of Data Correcting (DC) algorithms for the maximization of general submodular functions and its special cases, for example, the QCP and simple plant location problems (which is a special case of minimization a supermodular function). Cherenin's Excluding Rules, the dichotomy algorithm, and its generalization with the new branching rules are easy corollaries of our main result. The usefulness of our new branching rules is illustrated by means of a numerical Simple Plant Location Problem example.

http://www.springer.com/978-1-4614-5285-0

Data Correcting Approaches in Combinatorial Optimization

Goldengorin, B.; Pardalos, P.

2012, X, 114 p. 41 illus., Softcover
ISBN: 978-1-4614-5285-0