in a step-by-step fashion, one sees that it does better than any other algorithm at each step; it then follows that it produces an optimal solution. The second approach is known as an exchange argument, and it is more general: one considers any possible solution to the problem and gradually transforms it into the solution found by the greedy algorithm without altering its quality. Again, it will follow that the greedy algorithm must have found a solution that is at least as good as any other solution.

Following our introduction of these two styles of analysis, we focus on several of the most well-known applications of greedy algorithms: shortest paths in a graph, the Minimum Spanning Tree Problem, and the construction of Huffman codes for performing data compression. They each provide nice examples of our analysis techniques. We also explore an interesting relationship between minimum spanning trees and the long-studied problem of clustering. Finally, we consider a more complex application, the Minimum-Cost Arborescence Problem, which further extends our notion of what a greedy algorithm is.

4.1 Interval Scheduling: The Greedy Algorithm Stays Ahead

Let's recall the Interval Scheduling Problem, which was the first of the five representative problems we considered in Chapter 1. We have a set of requests \{1, 2, ..., n\}; the \(i\)th request corresponds to an interval of time starting at \(s(i)\) and finishing at \(f(i)\). (Note that we are slightly changing the notation from Section 1.2, where we used \(s_i\) rather than \(s(i)\) and \(f_i\) rather than \(f(i)\). This change of notation will make things easier to talk about in the proofs.) We'll say that a subset of the requests is compatible if no two of them overlap in time, and our goal is to accept as large a compatible subset as possible. Compatible sets of maximum size will be called optimal.

Designing a Greedy Algorithm

Using the Interval Scheduling Problem, we can make our discussion of greedy algorithms much more concrete. The basic idea in a greedy algorithm for interval scheduling is to use a simple rule to select a first request. Once a request is accepted, we reject all requests that are not compatible with it. We then select the next request to be accepted, and again reject all requests that are not compatible with it. We continue in this fashion until we run out of requests. The challenge in designing a good greedy algorithm is in deciding which simple rule to use for the selection—and there are many natural rules for this problem that do not give good solutions.

Let's try to think of some of the most natural rules and see how they work.

- The most obvious rule might be to always select the available request that starts earliest—that is, the one with minimal start time \(s(i)\). This way our resource starts being used as quickly as possible.
  
  This method does not yield an optimal solution. If the earliest request \(i\) is for a very long interval, then by accepting request \(i\) we may have to reject a lot of requests for shorter time intervals. Since our goal is to satisfy as many requests as possible, we will end up with a suboptimal solution.
  
  In a really bad case—say, when the finish time \(f(i)\) is the maximum among all requests—the accepted request \(i\) keeps our resource occupied for the whole time. In this case our greedy method would accept a single request, while the optimal solution could accept many. Such a situation is depicted in Figure 4.1(a).

- This might suggest that we should start out by accepting the request that requires the smallest interval of time—namely, the request for which \(f(i) - s(i)\) is as small as possible. As it turns out, this is a somewhat better rule than the previous one, but it still can produce a suboptimal schedule. For example, in Figure 4.1(b), accepting the short interval in the middle would prevent us from accepting the other two, which form an optimal solution.

Figure 4.1 Some instances of the Interval Scheduling Problem on which natural greedy algorithms fail to find the optimal solution. In (a), it does not work to select the interval that starts earliest; in (b), it does not work to select the shortest interval; and in (c), it does not work to select the interval with the fewest conflicts.
In the previous greedy rule, our problem was that the second request competes with both the first and the third—that is, accepting this request made us reject two other requests. We could design a greedy algorithm that is based on this idea: for each request, we count the number of other requests that are not compatible, and accept the request that has the fewest number of incompatible requests. (In other words, we select the interval with the fewest "conflicts.") This greedy choice would lead to the optimum solution in the previous example. In fact, it is quite a bit harder to design a bad example for this rule; but it can be done, and we've drawn an example in Figure 4.1(c). The unique optimal solution in this example is to accept the four requests in the top row. The greedy method suggested here accepts the middle request in the second row and thereby ensures a solution of size no greater than three.

A greedy rule that does lead to the optimal solution is based on a fourth idea: we should accept first the request that finishes first, that is, the request for which \( f(i) \) is as small as possible. This is also quite a natural idea: we ensure that our resource becomes free as soon as possible while still satisfying one request. In this way we can maximize the time left to satisfy other requests.

Let us state the algorithm a bit more formally. We will use \( R \) to denote the set of requests that we have neither accepted nor rejected yet, and use \( A \) to denote the set of accepted requests. For an example of how the algorithm runs, see Figure 4.2.

1. Initially let \( R \) be the set of all requests, and let \( A \) be empty.
2. While \( R \) is not yet empty:
   a. Choose a request \( i \) from \( R \) that has the smallest finishing time.
   b. Add request \( i \) to \( A \).
   c. Delete all requests from \( R \) that are not compatible with request \( i \).
3. Return the set \( A \) as the set of accepted requests.

Analyzing the Algorithm

While this greedy method is quite natural, it is certainly not obvious that it returns an optimal set of intervals. Indeed, it would only be sensible to reserve judgment on its optimality: the ideas that led to the previous nonoptimal versions of the greedy method also seemed promising at first.

As a start, we can immediately declare that the intervals in the set \( A \) returned by the algorithm are all compatible.

\[ (4.1) \quad A \text{ is a compatible set of requests.} \]
Our intuition for the greedy method came from wanting our resource to become free again as soon as possible after satisfying the first request. And indeed, our greedy rule guarantees that \( f(i_1) \leq f(i_2) \). This is the sense in which we want to show that our greedy rule “stays ahead”—that each of its intervals finishes no later than the corresponding interval in the set \( \mathcal{O} \). Thus we now prove that for each \( r \geq 1 \), the \( r \)-th interval in \( \mathcal{O} \) finishes at least as soon as the corresponding interval in the set \( \mathcal{O} \). Thus we now prove that for each \( r \geq 1 \), the \( r \)-th interval in \( \mathcal{O} \) finishes no later than the \( r \)-th interval in the optimal schedule.

\[(4.2) \quad \text{For all indices } r \leq k \text{ we have } f(i_r) \leq f(j_r). \]

**Proof.** We will prove this statement by induction. For \( r = 1 \) the statement is clearly true: the algorithm starts by selecting the request \( i_1 \) with minimum finish time.

Now let \( r > 1 \). We will assume as our induction hypothesis that the statement is true for \( r - 1 \), and we will try to prove it for \( r \). As shown in Figure 4.3, the induction hypothesis lets us assume that \( f(i_{r-1}) \leq f(j_{r-1}) \). In order for the algorithm’s \( r \)-th interval not to finish earlier as well, it would need to “fall behind” as shown. But there’s a simple reason why this could not happen: rather than choose a later-finishing interval, the greedy algorithm must make the same choice it made in the \( (r-1) \)-th step.

We can make this argument precise as follows. We know (since \( \mathcal{O} \) consists of compatible intervals) that \( f(i_{r-1}) \leq f(j_{r-1}) \). Combining this with the induction hypothesis \( f(i_{r-1}) \leq f(j_{r-1}) \), we get \( f(i_r) \leq f(j_r) \). Thus the interval \( j_r \) is in the set \( \mathcal{R} \) of available intervals at the time when the greedy algorithm selects \( i_r \). The greedy algorithm selects the available interval with smallest finish time; since interval \( j_r \) is one of these available intervals, we have \( f(i_r) \leq f(j_r) \). This completes the induction step.

Thus we have formalized the sense in which the greedy algorithm “stays ahead” of \( \mathcal{O} \): for each \( r \), the \( r \)-th interval it selects finishes at least as soon as the \( r \)-th interval in \( \mathcal{O} \). We now see why this implies the optimality of the greedy algorithm’s set \( \mathcal{A} \).

4.1 Interval Scheduling: The Greedy Algorithm Stays Ahead

**Proof.** We will prove the statement by contradiction. If \( \mathcal{A} \) is not optimal, then an optimal set \( \mathcal{O} \) must have more requests, that is, we must have \( m > k \). Applying (4.2) with \( r = k \), we get that \( f(i_k) \leq f(j_k) \). Since \( m > k \), there is a request \( j_{k+1} \) in \( \mathcal{O} \). This request starts after request \( j_k \) ends, and hence after \( i_k \) ends. So after deleting all requests that are not compatible with requests \( i_1, \ldots, i_k \), the set of possible requests \( \mathcal{R} \) still contains \( j_{k+1} \). But the greedy algorithm stops with request \( j_k \), and it is only supposed to stop when \( \mathcal{R} \) is empty—a contradiction.

**Implementation and Running Time** We can make our algorithm run in time \( O(n \log n) \) as follows. We begin by sorting the \( n \) requests in order of finishing time and labeling them in this order; that is, we will assume that \( f(i) \leq f(j) \) when \( i < j \). This takes time \( O(n \log n) \). In an additional \( O(n) \) time, we construct an array \( S[1 \ldots n] \) with the property that \( S[i] \) contains the value \( s(i) \).

We now select requests by processing the intervals in order of increasing \( f(i) \). We always select the first interval; then iterate through the intervals in order until reaching the first interval \( j \) for which \( s(i) \geq f(j) \); we then select this one as well. More generally, if the most recent interval we’ve selected ends at time \( f \), we continue iterating through subsequent intervals until we reach the first \( j \) for which \( s(i) \geq f \). In this way, we implement the greedy algorithm analyzed above in one pass through the intervals, spending constant time per interval. Thus this part of the algorithm takes time \( O(n) \).

**Extensions** The Interval Scheduling Problem we considered here is a quite simple scheduling problem. There are many further complications that could arise in practical settings. The following point out issues that we will see later in the book in various forms.

- In defining the problem, we assumed that all requests were known to the scheduling algorithm when it was choosing the compatible subset. It would also be natural, of course, to think about the version of the problem in which the scheduler needs to make decisions about accepting or rejecting certain requests before knowing about the full set of requests. Customers (requestors) may well be impatient, and they may give up and leave if the scheduler waits too long to gather information about all other requests. An active area of research is concerned with such online algorithms, which must make decisions as time proceeds, without knowledge of future input.
Our goal was to maximize the number of satisfied requests. But we could picture a situation in which each request has a different value to us. For example, each request could also have a value \( v_i \) (the amount gained by satisfying request \( i \)), and the goal would be to maximize our income: the sum of the values of all satisfied requests. This leads to the Weighted Interval Scheduling Problem, the second of the representative problems we described in Chapter 1.

There are many other variants and combinations that can arise. We now discuss one of these further variants in more detail, since it forms another case in which a greedy algorithm can be used to produce an optimal solution.

A Related Problem: Scheduling All Intervals

The Problem In the Interval Scheduling Problem, there is a single resource and many requests in the form of time intervals, so we must choose which requests to accept and which to reject. A related problem arises if we have many identical resources available and we wish to schedule all the requests using as few resources as possible. Because the goal here is to partition all intervals across multiple resources, we will refer to this as the Interval Partitioning Problem.\(^1\)

For example, suppose that each request corresponds to a lecture that needs to be scheduled in a classroom for a particular interval of time. We wish to satisfy all these requests, using as few classrooms as possible. The classrooms at our disposal are thus the multiple resources, and the basic constraint is that any two lectures that overlap in time must be scheduled in different classrooms. Equivalently, the interval requests could be jobs that need to be processed for a specific period of time, and the resources are machines capable of handling these jobs. Much later in the book, in Chapter 10, we will see a different application of this problem in which the intervals are routing requests that need to be allocated bandwidth on a fiber-optic cable.

As an illustration of the problem, consider the sample instance in Figure 4.4(a). The requests in this sample can all be scheduled using three resources; this is indicated in Figure 4.4(b), where the requests are rearranged into three rows, each containing a set of nonoverlapping intervals. In general, one can imagine a solution using \( k \) resources as a rearrangement of the requests into \( k \) rows of nonoverlapping intervals: the first row contains all the intervals assigned to the first resource, the second row contains all those assigned to the second resource, and so forth.

Now, is there any hope of using just two resources in this sample instance? Clearly the answer is no. We need at least three resources since, for example, intervals \( a, b, \) and \( c \) all pass over a common point on the time-line, and hence they all need to be scheduled on different resources. In fact, one can make this last argument in general for any instance of Interval Partitioning. Suppose we define the depth of a set of intervals to be the maximum number that pass over any single point on the time-line. Then we claim

\[ d \geq D \]

where \( D \) is the depth of the set of intervals.

Proof. Suppose a set of intervals has depth \( d \), and let \( I_1, \ldots, I_d \) all pass over a common point on the time-line. Then each of these intervals must be scheduled on a different resource, so the whole instance needs at least \( d \) resources.

We now consider two questions, which turn out to be closely related. First, can we design an efficient algorithm that schedules all intervals using the minimum possible number of resources? Second, is there always a schedule using a number of resources that is equal to the depth? In effect, a positive answer to this second question would say that the only obstacles to partitioning intervals are purely local—a set of intervals all piled over the same point. It's not immediately clear that there couldn't exist other, "long-range" obstacles that push the number of required resources even higher.

---

\(^1\) The problem is also referred to as the Interval Coloring Problem; the terminology arises from thinking of the different resources as having distinct colors—all the intervals assigned to a particular resource are given the corresponding color.
We now design a simple greedy algorithm that schedules all intervals using a number of resources equal to the depth. This immediately implies the optimality of the algorithm: in view of (4.4), no solution could use a number of resources that is smaller than the depth. The analysis of our algorithm will therefore illustrate another general approach to proving optimality: one finds a simple, "structural" bound asserting that every possible solution must have at least a certain value, and then one shows that the algorithm under consideration always achieves this bound.

**Designing the Algorithm**

Let \( d \) be the depth of the set of intervals; we show how to assign a label to each interval, where the labels come from the set of numbers \( \{1, 2, \ldots, d\} \), and the assignment has the property that overlapping intervals are labeled with different numbers. This gives the desired solution, since we can interpret each number as the name of a resource, and the label of each interval as the name of the resource to which it is assigned.

The algorithm we use for this is a simple one-pass greedy strategy that orders intervals by their starting times. We go through the intervals in this order, and try to assign to each interval we encounter a label that hasn't already been assigned to any previous interval that overlaps it. Specifically, we have the following description.

Sort the intervals by their start times, breaking ties arbitrarily

Let \( I_1, I_2, \ldots, I_n \) denote the intervals in this order

For \( j = 1, 2, 3, \ldots, n \)

For each interval \( I_i \) that precedes \( I_j \) in sorted order and overlaps it

Exclude the label of \( I_i \) from consideration for \( I_j \)

Endfor

If there is any label from \( \{1, 2, \ldots, d\} \) that has not been excluded then

Assign a nonexcluded label to \( I_j \)

Else

Leave \( I_j \) unlabeled

Endif

Endfor

**Analyzing the Algorithm**

We claim the following.

(4.5) If we use the greedy algorithm above, every interval will be assigned a label, and no two overlapping intervals will receive the same label.

**Proof.** First let's argue that no interval ends up unlabeled. Consider one of the intervals \( I_j \), and suppose there are \( t \) intervals earlier in the sorted order that overlap it. These \( t \) intervals, together with \( I_j \), form a set of \( t + 1 \) intervals that all pass over a common point on the time-line (namely, the start time of \( I_j \)), and so \( t + 1 \leq d \). Thus \( t \leq d - 1 \). It follows that at least one of the \( d \) labels is not excluded by this set of \( t \) intervals, and so there is a label that can be assigned to \( I_j \).

Next we claim that no two overlapping intervals are assigned the same label. Indeed, consider any two intervals \( I \) and \( I' \) that overlap, and suppose \( I \) precedes \( I' \) in the sorted order. Then when \( I' \) is considered by the algorithm, every resource name that it is assigned is in the set of resource names assigned to \( I \), and consequently the algorithm will not assign to \( I' \) the label that it used for \( I \). □

The algorithm and its analysis are very simple. Essentially, if you have \( d \) labels at your disposal, then as you sweep through the intervals from left to right, assigning an available label to each interval you encounter, you can never reach a point where all the labels are currently in use.

Since our algorithm is using \( d \) labels, we can use (4.4) to conclude that it is, in fact, always using the minimum possible number of labels. We sum this up as follows.

(4.6) The greedy algorithm above schedules every interval on a resource, using a number of resources equal to the depth of the set of intervals. This is the optimal number of resources needed.

**4.2 Scheduling to Minimize Lateness: An Exchange Argument**

We now discuss a scheduling problem related to the one with which we began the chapter. Despite the similarities in the problem formulation and in the greedy algorithm to solve it, the proof that this algorithm is optimal will require a more sophisticated kind of analysis.

**The Problem**

Consider again a situation in which we have a single resource and a set of \( n \) requests to use the resource for an interval of time. Assume that the resource is available starting at time \( s \). In contrast to the previous problem, however, each request is now more flexible. Instead of a start time and finish time, the request \( i \) has a deadline \( d_i \), and it requires a contiguous time interval of length \( t_i \), but it is willing to be scheduled at any time before the deadline. Each accepted request must be assigned an interval of time of length \( t_i \), and different requests must be assigned nonoverlapping intervals.

There are many objective functions we might seek to optimize when faced with this situation, and some are computationally much more difficult than...