CHAPTER 2

Algorithm Analysis

An **algorithm** is a clearly specified set of simple instructions to be followed to solve a problem. Once an algorithm is given for a problem and decided (somehow) to be correct, an important step is to determine how much in the way of resources, such as time or space, the algorithm will require. An algorithm that solves a problem but requires a year is hardly of any use. Likewise, an algorithm that requires thousands of gigabytes of main memory is not (currently) useful on most machines.

In this chapter, we shall discuss . . .

- How to estimate the time required for a program.
- How to reduce the running time of a program from days or years to fractions of a second.
- The results of careless use of recursion.
- Very efficient algorithms to raise a number to a power and to compute the greatest common divisor of two numbers.

2.1 Mathematical Background

The analysis required to estimate the resource use of an algorithm is generally a theoretical issue, and therefore a formal framework is required. We begin with some mathematical definitions.

Throughout this book, we will use the following four definitions:

Definition 2.1
$T(N) = O(f(N))$ if there are positive *constants* c and n_0 such that $T(N) \leq cf(N)$ when $N \geq n_0$.

Definition 2.2
$T(N) = \Omega(g(N))$ if there are positive *constants* c and n_0 such that $T(N) \geq cg(N)$ when $N \geq n_0$.

Definition 2.3
$T(N) = \Theta(h(N))$ if and only if $T(N) = O(h(N))$ and $T(N) = \Omega(h(N))$.

Definition 2.4
$T(N) = o(p(N))$ if, for all positive constants c, there exists an n_0 such that $T(N) < cp(N)$ when $N > n_0$. Less formally, $T(N) = o(p(N))$ if $T(N) = O(p(N))$ and $T(N) \neq \Theta(p(N))$.

The idea of these definitions is to establish a relative order among functions. Given two functions, there are usually points where one function is smaller than the other. So it does not make sense to claim, for instance, $f(N) < g(N)$. Thus, we compare their **relative rates of growth**. When we apply this to the analysis of algorithms, we shall see why this is the important measure.

Although $1{,}000N$ is larger than N^2 for small values of N, N^2 grows at a faster rate, and thus N^2 will eventually be the larger function. The turning point is $N = 1{,}000$ in this case. The first definition says that eventually there is some point n_0 past which $c \cdot f(N)$ is always at least as large as $T(N)$, so that if constant factors are ignored, $f(N)$ is at least as big as $T(N)$. In our case, we have $T(N) = 1{,}000N$, $f(N) = N^2$, $n_0 = 1{,}000$, and $c = 1$. We could also use $n_0 = 10$ and $c = 100$. Thus, we can say that $1{,}000N = O(N^2)$ (order N-squared). This notation is known as **Big-Oh notation**. Frequently, instead of saying "order . . . ," one says "Big-Oh . . . ."

If we use the traditional inequality operators to compare growth rates, then the first definition says that the growth rate of $T(N)$ is less than or equal to ($\leq$) that of $f(N)$. The second definition, $T(N) = \Omega(g(N))$ (pronounced "omega"), says that the growth rate of $T(N)$ is greater than or equal to ($\geq$) that of $g(N)$. The third definition, $T(N) = \Theta(h(N))$ (pronounced "theta"), says that the growth rate of $T(N)$ equals ($=$) the growth rate of $h(N)$. The last definition, $T(N) = o(p(N))$ (pronounced "little-oh"), says that the growth rate of $T(N)$ is less than ($<$) the growth rate of $p(N)$. This is different from Big-Oh, because Big-Oh allows the possibility that the growth rates are the same.

To prove that some function $T(N) = O(f(N))$, we usually do not apply these definitions formally but instead use a repertoire of known results. In general, this means that a proof (or determination that the assumption is incorrect) is a very simple calculation and should not involve calculus, except in extraordinary circumstances (not likely to occur in an algorithm analysis).

When we say that $T(N) = O(f(N))$, we are guaranteeing that the function $T(N)$ grows at a rate no faster than $f(N)$; thus $f(N)$ is an **upper bound** on $T(N)$. Since this implies that $f(N) = \Omega(T(N))$, we say that $T(N)$ is a **lower bound** on $f(N)$.

As an example, N^3 grows faster than N^2, so we can say that $N^2 = O(N^3)$ or $N^3 = \Omega(N^2)$. $f(N) = N^2$ and $g(N) = 2N^2$ grow at the same rate, so both $f(N) = O(g(N))$ and $f(N) = \Omega(g(N))$ are true. When two functions grow at the same rate, then the decision of whether or not to signify this with $\Theta()$ can depend on the particular context. Intuitively, if $g(N) = 2N^2$, then $g(N) = O(N^4)$, $g(N) = O(N^3)$, and $g(N) = O(N^2)$ are all technically correct, but the last option is the best answer. Writing $g(N) = \Theta(N^2)$ says not only that $g(N) = O(N^2)$ but also that the result is as good (tight) as possible.

Here are the important things to know:

Rule 1

If $T_1(N) = O(f(N))$ and $T_2(N) = O(g(N))$, then

(a) $T_1(N) + T_2(N) = O(f(N) + g(N))$ (intuitively and less formally it is $O(\max(f(N), g(N)))$),

(b) $T_1(N) * T_2(N) = O(f(N) * g(N))$.

Rule 2

If $T(N)$ is a polynomial of degree k, then $T(N) = \Theta(N^k)$.

Function	Name
c	Constant
$\log N$	Logarithmic
$\log^2 N$	Log-squared
N	Linear
$N \log N$	
N^2	Quadratic
N^3	Cubic
2^N	Exponential

Figure 2.1 Typical growth rates

Rule 3
$\log^k N = O(N)$ for any constant k. This tells us that logarithms grow very slowly.

This information is sufficient to arrange most of the common functions by growth rate (see Fig. 2.1).

Several points are in order. First, it is very bad style to include constants or low-order terms inside a Big-Oh. Do not say $T(N) = O(2N^2)$ or $T(N) = O(N^2 + N)$. In both cases, the correct form is $T(N) = O(N^2)$. This means that in any analysis that will require a Big-Oh answer, all sorts of shortcuts are possible. Lower-order terms can generally be ignored, and constants can be thrown away. Considerably less precision is required in these cases.

Second, we can always determine the relative growth rates of two functions $f(N)$ and $g(N)$ by computing $\lim_{N\to\infty} f(N)/g(N)$, using L'Hôpital's rule if necessary.[1] The limit can have four possible values:

- The limit is 0: This means that $f(N) = o(g(N))$.
- The limit is $c \neq 0$: This means that $f(N) = \Theta(g(N))$.
- The limit is ∞: This means that $g(N) = o(f(N))$.
- The limit does not exist: There is no relation (this will not happen in our context).

Using this method almost always amounts to overkill. Usually the relation between $f(N)$ and $g(N)$ can be derived by simple algebra. For instance, if $f(N) = N \log N$ and $g(N) = N^{1.5}$, then to decide which of $f(N)$ and $g(N)$ grows faster, one really needs to determine which of $\log N$ and $N^{0.5}$ grows faster. This is like determining which of $\log^2 N$ or N grows faster. This is a simple problem, because it is already known that N grows faster than any power of a log. Thus, $g(N)$ grows faster than $f(N)$.

One stylistic note: It is bad to say $f(N) \leq O(g(N))$, because the inequality is implied by the definition. It is wrong to write $f(N) \geq O(g(N))$, because it does not make sense.

[1] L'Hôpital's rule states that if $\lim_{N\to\infty} f(N) = \infty$ and $\lim_{N\to\infty} g(N) = \infty$, then $\lim_{N\to\infty} f(N)/g(N) = \lim_{N\to\infty} f'(N)/g'(N)$, where $f'(N)$ and $g'(N)$ are the derivatives of $f(N)$ and $g(N)$, respectively.

As an example of the typical kinds of analyses that are performed, consider the problem of downloading a file over the Internet. Suppose there is an initial 3-sec delay (to set up a connection), after which the download proceeds at 1.5M(bytes)/sec. Then it follows that if the file is N megabytes, the time to download is described by the formula $T(N) = N/1.5 + 3$. This is a **linear function**. Notice that the time to download a 1,500M file (1,003 sec) is approximately (but not exactly) twice the time to download a 750M file (503 sec). This is typical of a linear function. Notice, also, that if the speed of the connection doubles, both times decrease, but the 1,500M file still takes approximately twice the time to download as a 750M file. This is the typical characteristic of linear-time algorithms, and it is the reason we write $T(N) = O(N)$, ignoring constant factors. (Although using big-theta would be more precise, Big-Oh answers are typically given.)

Observe, too, that this behavior is not true of all algorithms. For the first selection algorithm described in Section 1.1, the running time is controlled by the time it takes to perform a sort. For a simple sorting algorithm, such as the suggested bubble sort, when the amount of input doubles, the running time increases by a factor of four for large amounts of input. This is because those algorithms are not linear. Instead, as we will see when we discuss sorting, trivial sorting algorithms are $O(N^2)$, or quadratic.

2.2 Model

In order to analyze algorithms in a formal framework, we need a model of computation. Our model is basically a normal computer in which instructions are executed sequentially. Our model has the standard repertoire of simple instructions, such as addition, multiplication, comparison, and assignment, but, unlike the case with real computers, it takes exactly one time unit to do anything (simple). To be reasonable, we will assume that, like a modern computer, our model has fixed-size (say, 32-bit) integers and no fancy operations, such as matrix inversion or sorting, which clearly cannot be done in one time unit. We also assume infinite memory.

This model clearly has some weaknesses. Obviously, in real life, not all operations take exactly the same time. In particular, in our model, one disk reads counts the same as an addition, even though the addition is typically several orders of magnitude faster. Also, by assuming infinite memory, we ignore the fact that the cost of a memory access can increase when slower memory is used due to larger memory requirements.

2.3 What to Analyze

The most important resource to analyze is generally the running time. Several factors affect the running time of a program. Some, such as the compiler and computer used, are obviously beyond the scope of any theoretical model, so, although they are important, we cannot deal with them here. The other main factors are the algorithm used and the input to the algorithm.

Typically, the size of the input is the main consideration. We define two functions, $T_{\text{avg}}(N)$ and $T_{\text{worst}}(N)$, as the average and worst-case running time, respectively, used by an algorithm on input of size N. Clearly, $T_{\text{avg}}(N) \leq T_{\text{worst}}(N)$. If there is more than one input, these functions may have more than one argument.

Occasionally, the best-case performance of an algorithm is analyzed. However, this is often of little interest, because it does not represent typical behavior. Average-case performance often reflects typical behavior, while worst-case performance represents a guarantee for performance on any possible input. Notice also that, although in this chapter we analyze C++ code, these bounds are really bounds for the algorithms rather than programs. Programs are an implementation of the algorithm in a particular programming language, and almost always the details of the programming language do not affect a Big-Oh answer. If a program is running much more slowly than the algorithm analysis suggests, there may be an implementation inefficiency. This can occur in C++ when arrays are inadvertently copied in their entirety, instead of passed with references. Another extremely subtle example of this is in the last two paragraphs of Section 12.6. Thus in future chapters, we will analyze the algorithms rather than the programs.

Generally, the quantity required is the worst-case time, unless otherwise specified. One reason for this is that it provides a bound for all input, including particularly bad input, which an average-case analysis does not provide. The other reason is that average-case bounds are usually much more difficult to compute. In some instances, the definition of "average" can affect the result. (For instance, what is average input for the following problem?)

As an example, in the next section, we shall consider the following problem:

Maximum Subsequence Sum Problem

Given (possibly negative) integers $A_1, A_2, \ldots, A_N$, find the maximum value of $\sum_{k=i}^{j} A_k$. (For convenience, the maximum subsequence sum is 0 if all the integers are negative.)

Example:

For input -2, 11, -4, 13, -5, -2, the answer is 20 (A_2 through A_4).

This problem is interesting mainly because there are so many algorithms to solve it, and the performance of these algorithms varies drastically. We will discuss four algorithms to solve this problem. The running time on some computers (the exact computer is unimportant) for these algorithms is given in Figure 2.2.

There are several important things worth noting in this table. For a small amount of input, the algorithms all run in the blink of an eye. So if only a small amount of input is

	Algorithm Time			
Input	1	2	3	4
Size	$O(N^3)$	$O(N^2)$	$O(N \log N)$	$O(N)$
$N = 100$	0.000159	0.000006	0.000005	0.000002
$N = 1{,}000$	0.095857	0.000371	0.000060	0.000022
$N = 10{,}000$	86.67	0.033322	0.000619	0.000222
$N = 100{,}000$	NA	3.33	0.006700	0.002205
$N = 1{,}000{,}000$	NA	NA	0.074870	0.022711

Figure 2.2 Running times of several algorithms for maximum subsequence sum (in seconds)

expected, it might be silly to expend a great deal of effort to design a clever algorithm. On the other hand, there is a large market these days for rewriting programs that were written five years ago based on a no-longer-valid assumption of small input size. These programs are now too slow because they used poor algorithms. For large amounts of input, algorithm 4 is clearly the best choice (although algorithm 3 is still usable).

Second, the times given do not include the time required to read the input. For algorithm 4, the time merely to read the input from a disk is likely to be an order of magnitude larger than the time required to solve the problem. This is typical of many efficient algorithms. Reading the data is generally the bottleneck; once the data are read, the problem can be solved quickly. For inefficient algorithms this is not true, and significant computer resources must be used. Thus, it is important that, whenever possible, algorithms be efficient enough not to be the bottleneck of a problem.

Notice that for algorithm 4, which is linear, as the problem size increases by a factor of 10, so does the running time. Algorithm 2, which is quadratic, does not display this behavior; a tenfold increase in input size yields roughly a hundredfold (10^2) increase in running time. And algorithm 1, which is cubic, yields a thousandfold (10^3) increase in running time. We would expect algorithm 1 to take nearly 9,000 seconds (or two and a half hours) to complete for $N = 100{,}000$. Similarly, we would expect algorithm 2 to take roughly 333 seconds to complete for $N = 1{,}000{,}000$. However, it is possible that algorithm 2 could take somewhat longer to complete due to the fact that $N = 1{,}000{,}000$ could also yield slower memory accesses than $N = 100{,}000$ on modern computers, depending on the size of the memory cache.

Figure 2.3 shows the growth rates of the running times of the four algorithms. Even though this graph encompasses only values of N ranging from 10 to 100, the relative

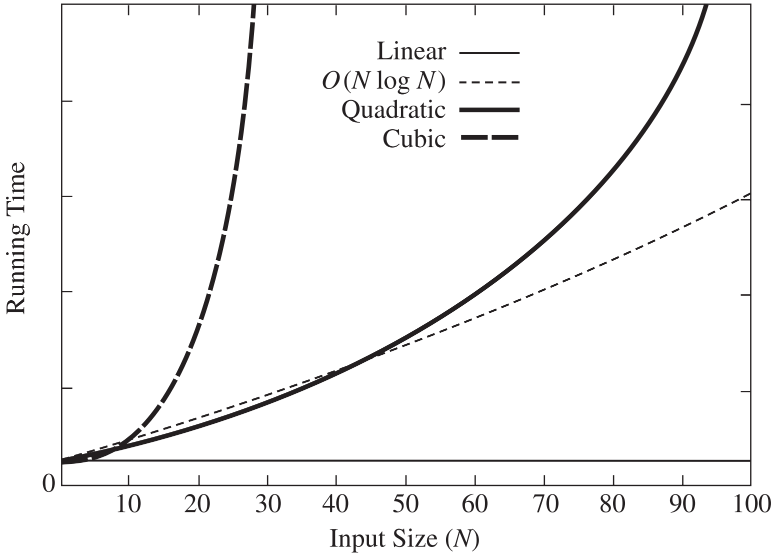

Figure 2.3 Plot (N vs. time) of various algorithms

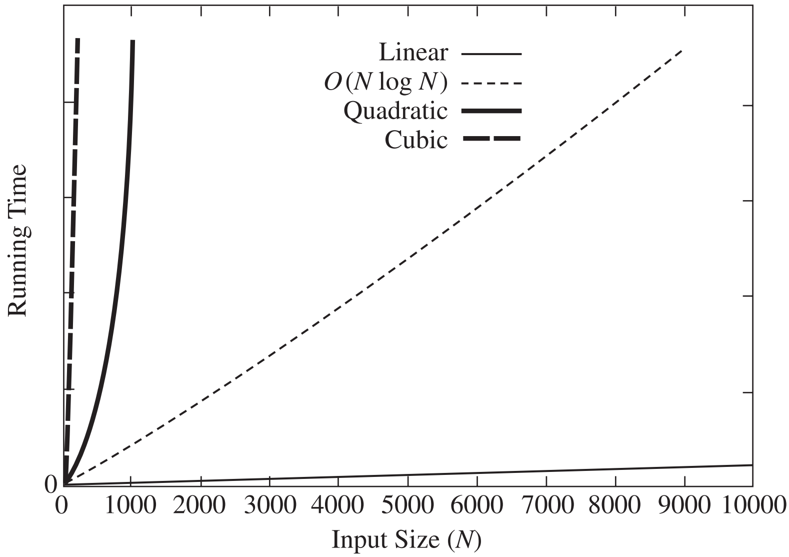

Figure 2.4 Plot (N vs. time) of various algorithms

growth rates are still evident. Although the graph for the $O(N \log N)$ seems linear, it is easy to verify that it is not by using a straightedge (or piece of paper). Although the graph for the $O(N)$ algorithm seems constant, this is only because for small values of N, the constant term is larger than the linear term. Figure 2.4 shows the performance for larger values. It dramatically illustrates how useless inefficient algorithms are for even moderately large amounts of input.

2.4 Running-Time Calculations

There are several ways to estimate the running time of a program. The previous table was obtained empirically. If two programs are expected to take similar times, probably the best way to decide which is faster is to code them both and run them!

Generally, there are several algorithmic ideas, and we would like to eliminate the bad ones early, so an analysis is usually required. Furthermore, the ability to do an analysis usually provides insight into designing efficient algorithms. The analysis also generally pinpoints the bottlenecks, which are worth coding carefully.

To simplify the analysis, we will adopt the convention that there are no particular units of time. Thus, we throw away leading constants. We will also throw away low-order terms, so what we are essentially doing is computing a Big-Oh running time. Since Big-Oh is an upper bound, we must be careful never to underestimate the running time of the program. In effect, the answer provided is a guarantee that the program will terminate within a certain time period. The program may stop earlier than this, but never later.

2.4.1 A Simple Example

Here is a simple program fragment to calculate $\sum_{i=1}^{N} i^3$:

```
        int sum( int n )
        {
            int partialSum;

1           partialSum = 0;
2           for( int i = 1; i <= n; ++i )
3               partialSum += i * i * i;
4           return partialSum;
        }
```

The analysis of this fragment is simple. The declarations count for no time. Lines 1 and 4 count for one unit each. Line 3 counts for four units per time executed (two multiplications, one addition, and one assignment) and is executed N times, for a total of $4N$ units. Line 2 has the hidden costs of initializing i, testing $i \leq N$, and incrementing i. The total cost of all these is 1 to initialize, $N + 1$ for all the tests, and N for all the increments, which is $2N + 2$. We ignore the costs of calling the function and returning, for a total of $6N + 4$. Thus, we say that this function is $O(N)$.

If we had to perform all this work every time we needed to analyze a program, the task would quickly become infeasible. Fortunately, since we are giving the answer in terms of Big-Oh, there are lots of shortcuts that can be taken without affecting the final answer. For instance, line 3 is obviously an $O(1)$ statement (per execution), so it is silly to count precisely whether it is two, three, or four units; it does not matter. Line 1 is obviously insignificant compared with the `for` loop, so it is silly to waste time here. This leads to several general rules.

2.4.2 General Rules

Rule 1—FOR loops
The running time of a `for` loop is at most the running time of the statements inside the `for` loop (including tests) times the number of iterations.

Rule 2—Nested loops
Analyze these inside out. The total running time of a statement inside a group of nested loops is the running time of the statement multiplied by the product of the sizes of all the loops.

As an example, the following program fragment is $O(N^2)$:

```
for( i = 0; i < n; ++i )
    for( j = 0; j < n; ++j )
        ++k;
```

Rule 3—Consecutive Statements
These just add (which means that the maximum is the one that counts; see rule 1 on page 52).

As an example, the following program fragment, which has $O(N)$ work followed by $O(N^2)$ work, is also $O(N^2)$:

```
for( i = 0; i < n; ++i )
    a[ i ] = 0;
for( i = 0; i < n; ++i )
    for( j = 0; j < n; ++j )
        a[ i ] += a[ j ] + i + j;
```

Rule 4—If/Else

For the fragment

```
if( condition )
    S1
else
    S2
```

the running time of an if/else statement is never more than the running time of the test plus the larger of the running times of S1 and S2.

Clearly, this can be an overestimate in some cases, but it is never an underestimate.

Other rules are obvious, but a basic strategy of analyzing from the inside (or deepest part) out works. If there are function calls, these must be analyzed first. If there are recursive functions, there are several options. If the recursion is really just a thinly veiled for loop, the analysis is usually trivial. For instance, the following function is really just a simple loop and is $O(N)$:

```
long factorial( int n )
{
    if( n <= 1 )
        return 1;
    else
        return n * factorial( n - 1 );
}
```

This example is really a poor use of recursion. When recursion is properly used, it is difficult to convert the recursion into a simple loop structure. In this case, the analysis will involve a recurrence relation that needs to be solved. To see what might happen, consider the following program, which turns out to be a terrible use of recursion:

```
long fib( int n )
{
    if( n <= 1 )
        return 1;
    else
       return fib( n - 1 ) + fib( n - 2 );
}
```

At first glance, this seems like a very clever use of recursion. However, if the program is coded up and run for values of N around 40, it becomes apparent that this program is terribly inefficient. The analysis is fairly simple. Let $T(N)$ be the running time for the function call `fib(n)`. If $N = 0$ or $N = 1$, then the running time is some constant value, which is the time to do the test at line 1 and return. We can say that $T(0) = T(1) = 1$ because constants do not matter. The running time for other values of N is then measured relative to the running time of the base case. For $N > 2$, the time to execute the function is the constant work at line 1 plus the work at line 3. Line 3 consists of an addition and two function calls. Since the function calls are not simple operations, they must be analyzed by themselves. The first function call is `fib(n-1)` and hence, by the definition of T, requires $T(N - 1)$ units of time. A similar argument shows that the second function call requires $T(N-2)$ units of time. The total time required is then $T(N-1)+T(N-2)+2$, where the 2 accounts for the work at line 1 plus the addition at line 3. Thus, for $N \geq 2$, we have the following formula for the running time of `fib(n)`:

$$T(N) = T(N - 1) + T(N - 2) + 2$$

Since `fib(n)` $=$ `fib(n-1)` $+$ `fib(n-2)`, it is easy to show by induction that $T(N) \geq$ `fib(n)`. In Section 1.2.5, we showed that $fib(N) < (5/3)^N$. A similar calculation shows that (for $N > 4$) $fib(N) \geq (3/2)^N$, and so the running time of this program grows *exponentially*. This is about as bad as possible. By keeping a simple array and using a `for` loop, the running time can be reduced substantially.

This program is slow because there is a huge amount of redundant work being performed, violating the fourth major rule of recursion (the compound interest rule), which was presented in Section 1.3. Notice that the first call on line 3, `fib(n-1)`, actually computes `fib(n-2)` at some point. This information is thrown away and recomputed by the second call on line 3. The amount of information thrown away compounds recursively and results in the huge running time. This is perhaps the finest example of the maxim "Don't compute anything more than once" and should not scare you away from using recursion. Throughout this book, we shall see outstanding uses of recursion.

2.4.3 Solutions for the Maximum Subsequence Sum Problem

We will now present four algorithms to solve the maximum subsequence sum problem posed earlier. The first algorithm, which merely exhaustively tries all possibilities, is depicted in Figure 2.5. The indices in the `for` loop reflect the fact that in C++, arrays begin at 0 instead of 1. Also, the algorithm does not compute the actual subsequences; additional code is required to do this.

Convince yourself that this algorithm works (this should not take much convincing). The running time is $O(N^3)$ and is entirely due to lines 13 and 14, which consist of an $O(1)$ statement buried inside three nested `for` loops. The loop at line 8 is of size N.

The second loop has size $N - i$, which could be small but could also be of size N. We must assume the worst, with the knowledge that this could make the final bound a bit high. The third loop has size $j - i + 1$, which again we must assume is of size N. The total is $O(1 \cdot N \cdot N \cdot N) = O(N^3)$. Line 6 takes only $O(1)$ total, and lines 16 and 17 take only $O(N^2)$ total, since they are easy expressions inside only two loops.

```
 1  /**
 2   * Cubic maximum contiguous subsequence sum algorithm.
 3   */
 4  int maxSubSum1( const vector<int> & a )
 5  {
 6      int maxSum = 0;
 7
 8      for( int i = 0; i < a.size( ); ++i )
 9          for( int j = i; j < a.size( ); ++j )
10          {
11              int thisSum = 0;
12
13              for( int k = i; k <= j; ++k )
14                  thisSum += a[ k ];
15
16              if( thisSum > maxSum )
17                  maxSum = thisSum;
18          }
19
20      return maxSum;
21  }
```

Figure 2.5 Algorithm 1

It turns out that a more precise analysis, taking into account the actual size of these loops, shows that the answer is $\Theta(N^3)$ and that our estimate above was a factor of 6 too high (which is all right, because constants do not matter). This is generally true in these kinds of problems. The precise analysis is obtained from the sum $\sum_{i=0}^{N-1} \sum_{j=i}^{N-1} \sum_{k=i}^{j} 1$, which tells how many times line 14 is executed. The sum can be evaluated inside out, using formulas from Section 1.2.3. In particular, we will use the formulas for the sum of the first N integers and first N squares. First we have

$$\sum_{k=i}^{j} 1 = j - i + 1$$

Next we evaluate

$$\sum_{j=i}^{N-1} (j - i + 1) = \frac{(N - i + 1)(N - i)}{2}$$

This sum is computed by observing that it is just the sum of the first $N - i$ integers. To complete the calculation, we evaluate

$$\begin{aligned}
\sum_{i=0}^{N-1} \frac{(N-i+1)(N-i)}{2} &= \sum_{i=1}^{N} \frac{(N-i+1)(N-i+2)}{2} \\
&= \frac{1}{2}\sum_{i=1}^{N} i^2 - \left(N + \frac{3}{2}\right)\sum_{i=1}^{N} i + \frac{1}{2}(N^2 + 3N + 2)\sum_{i=1}^{N} 1 \\
&= \frac{1}{2}\frac{N(N+1)(2N+1)}{6} - \left(N + \frac{3}{2}\right)\frac{N(N+1)}{2} + \frac{N^2 + 3N + 2}{2}N \\
&= \frac{N^3 + 3N^2 + 2N}{6}
\end{aligned}$$

We can avoid the cubic running time by removing a `for` loop. This is not always possible, but in this case there are an awful lot of unnecessary computations present in the algorithm. The inefficiency that the improved algorithm corrects can be seen by noticing that $\sum_{k=i}^{j} A_k = A_j + \sum_{k=i}^{j-1} A_k$, so the computation at lines 13 and 14 in algorithm 1 is unduly expensive. Figure 2.6 shows an improved algorithm. Algorithm 2 is clearly $O(N^2)$; the analysis is even simpler than before.

There is a recursive and relatively complicated $O(N \log N)$ solution to this problem, which we now describe. If there didn't happen to be an $O(N)$ (linear) solution, this would be an excellent example of the power of recursion. The algorithm uses a "divide-and-conquer" strategy. The idea is to split the problem into two roughly equal subproblems,

```
/**
 * Quadratic maximum contiguous subsequence sum algorithm.
 */
int maxSubSum2( const vector<int> & a )
{
    int maxSum = 0;

    for( int i = 0; i < a.size( ); ++i )
    {
        int thisSum = 0;
        for( int j = i; j < a.size( ); ++j )
        {
            thisSum += a[ j ];

            if( thisSum > maxSum )
                maxSum = thisSum;
        }
    }

    return maxSum;
}
```

Figure 2.6 Algorithm 2

which are then solved recursively. This is the "divide" part. The "conquer" stage consists of patching together the two solutions of the subproblems, and possibly doing a small amount of additional work, to arrive at a solution for the whole problem.

In our case, the maximum subsequence sum can be in one of three places. Either it occurs entirely in the left half of the input, or entirely in the right half, or it crosses the middle and is in both halves. The first two cases can be solved recursively. The last case can be obtained by finding the largest sum in the first half that includes the last element in the first half, and the largest sum in the second half that includes the first element in the second half. These two sums can then be added together. As an example, consider the following input:

First Half				Second Half			
4	−3	5	−2	−1	2	6	−2

The maximum subsequence sum for the first half is 6 (elements A_1 through A_3) and for the second half is 8 (elements A_6 through A_7).

The maximum sum in the first half that includes the last element in the first half is 4 (elements A_1 through A_4), and the maximum sum in the second half that includes the first element in the second half is 7 (elements A_5 through A_7). Thus, the maximum sum that spans both halves and goes through the middle is $4 + 7 = 11$ (elements A_1 through A_7).

We see, then, that among the three ways to form a large maximum subsequence, for our example, the best way is to include elements from both halves. Thus, the answer is 11. Figure 2.7 shows an implementation of this strategy.

The code for algorithm 3 deserves some comment. The general form of the call for the recursive function is to pass the input array along with the left and right borders, which delimits the portion of the array that is operated upon. A one-line driver program sets this up by passing the borders 0 and $N - 1$ along with the array.

Lines 8 to 12 handle the base case. If `left == right`, there is one element, and it is the maximum subsequence if the element is nonnegative. The case `left > right` is not possible unless N is negative (although minor perturbations in the code could mess this up). Lines 15 and 16 perform the two recursive calls. We can see that the recursive calls are always on a smaller problem than the original, although minor perturbations in the code could destroy this property. Lines 18 to 24 and 26 to 32 calculate the two maximum sums that touch the center divider. The sum of these two values is the maximum sum that spans both halves. The routine `max3` (not shown) returns the largest of the three possibilities.

Algorithm 3 clearly requires more effort to code than either of the two previous algorithms. However, shorter code does not always mean better code. As we have seen in the earlier table showing the running times of the algorithms, this algorithm is considerably faster than the other two for all but the smallest of input sizes.

The running time is analyzed in much the same way as for the program that computes the Fibonacci numbers. Let $T(N)$ be the time it takes to solve a maximum subsequence sum problem of size N. If $N = 1$, then the program takes some constant amount of time to execute lines 8 to 12, which we shall call one unit. Thus, $T(1) = 1$. Otherwise, the

```
/**
 * Recursive maximum contiguous subsequence sum algorithm.
 * Finds maximum sum in subarray spanning a[left..right].
 * Does not attempt to maintain actual best sequence.
 */
int maxSumRec( const vector<int> & a, int left, int right )
{
    if( left == right )  // Base case
        if( a[ left ] > 0 )
            return a[ left ];
        else
            return 0;

    int center = ( left + right ) / 2;
    int maxLeftSum  = maxSumRec( a, left, center );
    int maxRightSum = maxSumRec( a, center + 1, right );

    int maxLeftBorderSum = 0, leftBorderSum = 0;
    for( int i = center; i >= left; --i )
    {
        leftBorderSum += a[ i ];
        if( leftBorderSum > maxLeftBorderSum )
            maxLeftBorderSum = leftBorderSum;
    }

    int maxRightBorderSum = 0, rightBorderSum = 0;
    for( int j = center + 1; j <= right; ++j )
    {
        rightBorderSum += a[ j ];
        if( rightBorderSum > maxRightBorderSum )
            maxRightBorderSum = rightBorderSum;
    }

    return max3( maxLeftSum, maxRightSum,
                    maxLeftBorderSum + maxRightBorderSum );
}

/**
 * Driver for divide-and-conquer maximum contiguous
 * subsequence sum algorithm.
 */
int maxSubSum3( const vector<int> & a )
{
    return maxSumRec( a, 0, a.size( ) - 1 );
}
```

Figure 2.7 Algorithm 3

program must perform two recursive calls, the two for loops between lines 19 and 32, and some small amount of bookkeeping, such as lines 14 and 34. The two for loops combine to touch every element in the subarray, and there is constant work inside the loops, so the time expended in lines 19 to 32 is $O(N)$. The code in lines 8 to 14, 18, 26, and 34 is all a constant amount of work and can thus be ignored compared with $O(N)$. The remainder of the work is performed in lines 15 and 16. These lines solve two subsequence problems of size $N/2$ (assuming N is even). Thus, these lines take $T(N/2)$ units of time each, for a total of $2T(N/2)$. The total time for the algorithm then is $2T(N/2) + O(N)$. This gives the equations

$$T(1) = 1$$
$$T(N) = 2T(N/2) + O(N)$$

To simplify the calculations, we can replace the $O(N)$ term in the equation above with N; since $T(N)$ will be expressed in Big-Oh notation anyway, this will not affect the answer. In Chapter 7, we shall see how to solve this equation rigorously. For now, if $T(N) = 2T(N/2)+N$, and $T(1) = 1$, then $T(2) = 4 = 2*2$, $T(4) = 12 = 4*3$, $T(8) = 32 = 8*4$, and $T(16) = 80 = 16*5$. The pattern that is evident, and can be derived, is that if $N = 2^k$, then $T(N) = N * (k + 1) = N \log N + N = O(N \log N)$.

This analysis assumes N is even, since otherwise $N/2$ is not defined. By the recursive nature of the analysis, it is really valid only when N is a power of 2, since otherwise we eventually get a subproblem that is not an even size, and the equation is invalid. When N is not a power of 2, a somewhat more complicated analysis is required, but the Big-Oh result remains unchanged.

In future chapters, we will see several clever applications of recursion. Here, we present a fourth algorithm to find the maximum subsequence sum. This algorithm is simpler to implement than the recursive algorithm and also is more efficient. It is shown in Figure 2.8.

It should be clear why the time bound is correct, but it takes a little thought to see why the algorithm actually works. To sketch the logic, note that like algorithms 1 and 2, j is representing the end of the current sequence, while i is representing the start of the current sequence. It happens that the use of i can be optimized out of the program if we do not need to know where the actual best subsequence is, but in designing the algorithm, let's pretend that i is needed and that we are trying to improve algorithm 2. One observation is that if a[i] is negative, then it cannot possibly be the start of the optimal subsequence, since any subsequence that begins by including a[i] would be improved by beginning with a[i+1]. Similarly, any negative subsequence cannot possibly be a prefix of the optimal subsequence (same logic). If, in the inner loop, we detect that the subsequence from a[i] to a[j] is negative, then we can advance i. The crucial observation is that not only can we advance i to i+1, but we can also actually advance it all the way to j+1. To see this, let p be any index between i+1 and j. Any subsequence that starts at index p is not larger than the corresponding subsequence that starts at index i and includes the subsequence from a[i] to a[p-1], since the latter subsequence is not negative (j is the first index that causes the subsequence starting at index i to become negative). Thus, advancing i to j+1 is risk free; we cannot miss an optimal solution.

This algorithm is typical of many clever algorithms: The running time is obvious, but the correctness is not. For these algorithms, formal correctness proofs (more formal

```
/**
 * Linear-time maximum contiguous subsequence sum algorithm.
 */
int maxSubSum4( const vector<int> & a )
{
    int maxSum = 0, thisSum = 0;

    for( int j = 0; j < a.size( ); ++j )
    {
        thisSum += a[ j ];

        if( thisSum > maxSum )
            maxSum = thisSum;
        else if( thisSum < 0 )
            thisSum = 0;
    }

    return maxSum;
}
```

Figure 2.8 Algorithm 4

than the sketch above) are almost always required; even then, many people still are not convinced. Also, many of these algorithms require trickier programming, leading to longer development. But when these algorithms work, they run quickly, and we can test much of the code logic by comparing it with an inefficient (but easily implemented) brute-force algorithm using small input sizes.

An extra advantage of this algorithm is that it makes only one pass through the data, and once `a[i]` is read and processed, it does not need to be remembered. Thus, if the array is on a disk or is being transmitted over the Internet, it can be read sequentially, and there is no need to store any part of it in main memory. Furthermore, at any point in time, the algorithm can correctly give an answer to the subsequence problem for the data it has already read (the other algorithms do not share this property). Algorithms that can do this are called **online algorithms**. An online algorithm that requires only constant space and runs in linear time is just about as good as possible.

2.4.4 Logarithms in the Running Time

The most confusing aspect of analyzing algorithms probably centers around the logarithm. We have already seen that some divide-and-conquer algorithms will run in $O(N \log N)$ time. Besides divide-and-conquer algorithms, the most frequent appearance of logarithms centers around the following general rule: *An algorithm is $O(\log N)$ if it takes constant ($O(1)$) time to cut the problem size by a fraction (which is usually $\frac{1}{2}$).* On the other hand, if constant time is required to merely reduce the problem by a constant *amount* (such as to make the problem smaller by 1), then the algorithm is $O(N)$.

It should be obvious that only special kinds of problems can be $O(\log N)$. For instance, if the input is a list of N numbers, an algorithm must take $\Omega(N)$ merely to read the input in. Thus, when we talk about $O(\log N)$ algorithms for these kinds of problems, we usually presume that the input is preread. We provide three examples of logarithmic behavior.

Binary Search

The first example is usually referred to as binary search.

> **Binary Search**
> Given an integer X and integers $A_0, A_1, \ldots, A_{N-1}$, which are presorted and already in memory, find i such that $A_i = X$, or return $i = -1$ if X is not in the input.

The obvious solution consists of scanning through the list from left to right and runs in linear time. However, this algorithm does not take advantage of the fact that the list is sorted and is thus not likely to be best. A better strategy is to check if X is the middle element. If so, the answer is at hand. If X is smaller than the middle element, we can apply the same strategy to the sorted subarray to the left of the middle element; likewise, if X is larger than the middle element, we look to the right half. (There is also the case of when to stop.) Figure 2.9 shows the code for binary search (the answer is `mid`). As usual, the code reflects C++'s convention that arrays begin with index 0.

```
/**
 * Performs the standard binary search using two comparisons per level.
 * Returns index where item is found or -1 if not found.
 */
template <typename Comparable>
int binarySearch( const vector<Comparable> & a, const Comparable & x )
{
    int low = 0, high = a.size( ) - 1;

    while( low <= high )
    {
        int mid = ( low + high ) / 2;

        if( a[ mid ] < x )
            low = mid + 1;
        else if( a[ mid ] > x )
            high = mid - 1;
        else
            return mid;   // Found
    }
    return NOT_FOUND;     // NOT_FOUND is defined as -1
}
```

Figure 2.9 Binary search

Clearly, all the work done inside the loop takes $O(1)$ per iteration, so the analysis requires determining the number of times around the loop. The loop starts with `high - low` $= N - 1$ and finishes with `high - low` ≥ -1. Every time through the loop, the value `high - low` must be at least halved from its previous value; thus, the number of times around the loop is at most $\lceil \log(N-1) \rceil + 2$. (As an example, if `high - low` $= 128$, then the maximum values of `high - low` after each iteration are 64, 32, 16, 8, 4, 2, 1, 0, −1.) Thus, the running time is $O(\log N)$. Equivalently, we could write a recursive formula for the running time, but this kind of brute-force approach is usually unnecessary when you understand what is really going on and why.

Binary search can be viewed as our first data-structure implementation. It supports the `contains` operation in $O(\log N)$ time, but all other operations (in particular, `insert`) require $O(N)$ time. In applications where the data are static (i.e., insertions and deletions are not allowed), this could be very useful. The input would then need to be sorted once, but afterward accesses would be fast. An example is a program that needs to maintain information about the periodic table of elements (which arises in chemistry and physics). This table is relatively stable, as new elements are added infrequently. The element names could be kept sorted. Since there are only about 118 elements, at most eight accesses would be required to find an element. Performing a sequential search would require many more accesses.

Euclid's Algorithm

A second example is Euclid's algorithm for computing the greatest common divisor. The greatest common divisor (*gcd*) of two integers is the largest integer that divides both. Thus, $gcd(50, 15) = 5$. The algorithm in Figure 2.10 computes $gcd(M, N)$, assuming $M \geq N$. (If $N > M$, the first iteration of the loop swaps them.)

The algorithm works by continually computing remainders until 0 is reached. The last nonzero remainder is the answer. Thus, if $M = 1{,}989$ and $N = 1{,}590$, then the sequence of remainders is 399, 393, 6, 3, 0. Therefore, $gcd(1989, 1590) = 3$. As the example shows, this is a fast algorithm.

As before, estimating the entire running time of the algorithm depends on determining how long the sequence of remainders is. Although $\log N$ seems like a good answer, it is not at all obvious that the value of the remainder has to decrease by a constant factor,

```
long long gcd( long long m, long long n )
{
    while( n != 0 )
    {
        long long rem = m % n;
        m = n;
        n = rem;
    }
    return m;
}
```

Figure 2.10 Euclid's algorithm

since we see that the remainder went from 399 to only 393 in the example. Indeed, the remainder *does not* decrease by a constant factor in one iteration. However, we can prove that after two iterations, the remainder is at most half of its original value. This would show that the number of iterations is at most $2 \log N = O(\log N)$ and establish the running time. This proof is easy, so we include it here. It follows directly from the following theorem.

Theorem 2.1

If $M > N$, then $M \bmod N < M/2$.

Proof

There are two cases. If $N \leq M/2$, then since the remainder is smaller than N, the theorem is true for this case. The other case is $N > M/2$. But then N goes into M once with a remainder $M - N < M/2$, proving the theorem.

One might wonder if this is the best bound possible, since $2 \log N$ is about 20 for our example, and only seven operations were performed. It turns out that the constant can be improved slightly, to roughly $1.44 \log N$, in the worst case (which is achievable if M and N are consecutive Fibonacci numbers). The average-case performance of Euclid's algorithm requires pages and pages of highly sophisticated mathematical analysis, and it turns out that the average number of iterations is about $(12 \ln 2 \ln N)/\pi^2 + 1.47$.

Exponentiation

Our last example in this section deals with raising an integer to a power (which is also an integer). Numbers that result from exponentiation are generally quite large, so an analysis works only if we can assume that we have a machine that can store such large integers (or a compiler that can simulate this). We will count the number of multiplications as the measurement of running time.

The obvious algorithm to compute X^N uses $N-1$ multiplications. A recursive algorithm can do better. $N \leq 1$ is the base case of the recursion. Otherwise, if N is even, we have $X^N = X^{N/2} \cdot X^{N/2}$, and if N is odd, $X^N = X^{(N-1)/2} \cdot X^{(N-1)/2} \cdot X$.

For instance, to compute X^{62}, the algorithm does the following calculations, which involve only nine multiplications:

$$X^3 = (X^2)X, X^7 = (X^3)^2X, X^{15} = (X^7)^2X, X^{31} = (X^{15})^2X, X^{62} = (X^{31})^2$$

The number of multiplications required is clearly at most $2 \log N$, because at most two multiplications (if N is odd) are required to halve the problem. Again, a recurrence formula can be written and solved. Simple intuition obviates the need for a brute-force approach.

Figure 2.11 implements this idea. It is sometimes interesting to see how much the code can be tweaked without affecting correctness. In Figure 2.11, lines 5 to 6 are actually unnecessary, because if N is 1, then line 10 does the right thing. Line 10 can also be rewritten as

```
10          return pow( x, n - 1 ) * x;
```

```
1   long long pow( long-long x, int n )
2   {
3       if( n == 0 )
4           return 1;
5       if( n == 1 )
6           return x;
7       if( isEven( n ) )
8           return pow( x * x, n / 2 );
9       else
10          return pow( x * x, n / 2 ) * x;
11  }
```

Figure 2.11 Efficient exponentiation

without affecting the correctness of the program. Indeed, the program will still run in $O(\log N)$, because the sequence of multiplications is the same as before. However, all of the following alternatives for line 8 are bad, even though they look correct:

```
8a          return pow( pow( x, 2 ), n / 2 );
8b          return pow( pow( x, n / 2 ), 2 );
8c          return pow( x, n / 2 ) * pow( x, n / 2 );
```

Both lines 8a and 8b are incorrect because when N is 2, one of the recursive calls to `pow` has 2 as the second argument. Thus no progress is made, and an infinite loop results (in an eventual crash).

Using line 8c affects the efficiency, because there are now two recursive calls of size $N/2$ instead of only one. An analysis will show that the running time is no longer $O(\log N)$. We leave it as an exercise to the reader to determine the new running time.

2.4.5 Limitations of Worst-Case Analysis

Sometimes the analysis is shown empirically to be an overestimate. If this is the case, then either the analysis needs to be tightened (usually by a clever observation), or it may be that the *average* running time is significantly less than the worst-case running time and no improvement in the bound is possible. For many complicated algorithms the worst-case bound is achievable by some bad input but is usually an overestimate in practice. Unfortunately, for most of these problems, an average-case analysis is extremely complex (in many cases still unsolved), and a worst-case bound, even though overly pessimistic, is the best analytical result known.

Summary

This chapter gives some hints on how to analyze the complexity of programs. Unfortunately, it is not a complete guide. Simple programs usually have simple analyses, but this is not always the case. As an example, later in the text we shall see a sorting algorithm (Shellsort, Chapter 7) and an algorithm for maintaining disjoint sets (Chapter 8), each of which requires about 20 lines of code. The analysis of Shellsort is still not complete, and the disjoint set algorithm has an analysis that until recently was extremely difficult and require pages and pages of intricate calculations. Most of the analyses that we will encounter here will be simple and involve counting through loops.

An interesting kind of analysis, which we have not touched upon, is lower-bound analysis. We will see an example of this in Chapter 7, where it is proved that any algorithm that sorts by using only comparisons requires $\Omega(N \log N)$ comparisons in the worst case. Lower-bound proofs are generally the most difficult, because they apply not to an algorithm but to a class of algorithms that solve a problem.

We close by mentioning that some of the algorithms described here have real-life application. The *gcd* algorithm and the exponentiation algorithm are both used in cryptography. Specifically, a 600-digit number is raised to a large power (usually another 600-digit number), with only the low 600 or so digits retained after each multiplication. Since the calculations require dealing with 600-digit numbers, efficiency is obviously important. The straightforward algorithm for exponentiation would require about 10^{600} multiplications, whereas the algorithm presented requires only about 4,000 in the worst case.

Exercises

2.1 Order the following functions by growth rate: N, $\sqrt{N}$, $N^{1.5}$, N^2, $N \log N$, $N \log \log N$, $N \log^2 N$, $N \log(N^2)$, $2/N$, 2^N, $2^{N/2}$, 37, $N^2 \log N$, N^3. Indicate which functions grow at the same rate.

2.2 Suppose $T_1(N) = O(f(N))$ and $T_2(N) = O(f(N))$. Which of the following are true?
a. $T_1(N) + T_2(N) = O(f(N))$
b. $T_1(N) - T_2(N) = o(f(N))$
c. $\dfrac{T_1(N)}{T_2(N)} = O(1)$
d. $T_1(N) = O(T_2(N))$

2.3 Which function grows faster: $N \log N$ or $N^{1+\epsilon/\sqrt{\log N}}$, $\epsilon > 0$?

2.4 Prove that for any constant k, $\log^k N = o(N)$.

2.5 Find two functions $f(N)$ and $g(N)$ such that neither $f(N) = O(g(N))$ nor $g(N) = O(f(N))$.

2.6 In a recent court case, a judge cited a city for contempt and ordered a fine of \$2 for the first day. Each subsequent day, until the city followed the judge's order, the fine was squared (i.e., the fine progressed as follows: \$2, \$4, \$16, \$256, \$65,536, . . .).
a. What would be the fine on day N?
b. How many days would it take for the fine to reach D dollars (a Big-Oh answer will do)?

2.7 For each of the following six program fragments:
a. Give an analysis of the running time (Big-Oh will do).
b. Implement the code in the language of your choice, and give the running time for several values of N.
c. Compare your analysis with the actual running times.

```
(1)  sum = 0;
     for( i = 0; i < n; ++i )
         ++sum;
(2)  sum = 0;
     for( i = 0; i < n; ++i )
         for( j = 0; j < n; ++j )
             ++sum;
(3)  sum = 0;
     for( i = 0; i < n; ++i )
         for( j = 0; j < n * n; ++j )
             ++sum;
(4)  sum = 0;
     for( i = 0; i < n; ++i )
         for( j = 0; j < i; ++j )
             ++sum;
(5)  sum = 0;
     for( i = 0; i < n; ++i )
         for( j = 0; j < i * i; ++j )
             for( k = 0; k < j; ++k )
                 ++sum;
(6)  sum = 0;
     for( i = 1; i < n; ++i )
         for( j = 1; j < i * i; ++j )
             if( j % i == 0 )
                 for( k = 0; k < j; ++k )
                     ++sum;
```

2.8 Suppose you need to generate a *random* permutation of the first N integers. For example, {4, 3, 1, 5, 2} and {3, 1, 4, 2, 5} are legal permutations, but {5, 4, 1, 2, 1} is not, because one number (1) is duplicated and another (3) is missing. This routine is often used in simulation of algorithms. We assume the existence of a random number generator, `r`, with method `randInt(i,j)`, that generates integers between `i` and `j` with equal probability. Here are three algorithms:

1. Fill the array `a` from `a[0]` to `a[N-1]` as follows: To fill `a[i]`, generate random numbers until you get one that is not already in `a[0]`, `a[1]`, . . . , `a[i-1]`.
2. Same as algorithm (1), but keep an extra array called the `used` array. When a random number, `ran`, is first put in the array `a`, set `used[ran] = true`. This means that when filling `a[i]` with a random number, you can test in one step to see whether the random number has been used, instead of the (possibly) `i` steps in the first algorithm.
3. Fill the array such that `a[i] = i+1`. Then

```
for( i = 1; i < n; ++i )
    swap( a[ i ], a[ randInt( 0, i ) ] );
```

a. Prove that all three algorithms generate only legal permutations and that all permutations are equally likely.

b. Give as accurate (Big-Oh) an analysis as you can of the *expected* running time of each algorithm.
c. Write (separate) programs to execute each algorithm 10 times, to get a good average. Run program (1) for $N = 250, 500, 1{,}000, 2{,}000$; program (2) for $N = 25{,}000, 50{,}000, 100{,}000, 200{,}000, 400{,}000, 800{,}000$; and program (3) for $N = 100{,}000, 200{,}000, 400{,}000, 800{,}000, 1{,}600{,}000, 3{,}200{,}000, 6{,}400{,}000$.
d. Compare your analysis with the actual running times.
e. What is the worst-case running time of each algorithm?

2.9 Complete the table in Figure 2.2 with estimates for the running times that were too long to simulate. Interpolate the running times for these algorithms and estimate the time required to compute the maximum subsequence sum of 1 million numbers. What assumptions have you made?

2.10 Determine, for the typical algorithms that you use to perform calculations by hand, the running time to do the following:
a. Add two N-digit integers.
b. Multiply two N-digit integers.
c. Divide two N-digit integers.

2.11 An algorithm takes 0.5 ms for input size 100. How long will it take for input size 500 if the running time is the following (assume low-order terms are negligible)?
a. linear
b. $O(N \log N)$
c. quadratic
d. cubic

2.12 An algorithm takes 0.5 ms for input size 100. How large a problem can be solved in 1 min if the running time is the following (assume low-order terms are negligible)?
a. linear
b. $O(N \log N)$
c. quadratic
d. cubic

2.13 How much time is required to compute $f(x) = \sum_{i=0}^{N} a_i x^i$:
a. Using a simple routine to perform exponentiation?
b. Using the routine in Section 2.4.4?

2.14 Consider the following algorithm (known as *Horner's rule*) to evaluate $f(x) = \sum_{i=0}^{N} a_i x^i$:

```
poly = 0;
for( i = n; i >= 0; --i )
    poly = x * poly + a[i];
```

a. Show how the steps are performed by this algorithm for $x = 3$, $f(x) = 4x^4 + 8x^3 + x + 2$.
b. Explain why this algorithm works.
c. What is the running time of this algorithm?

2.15 Give an efficient algorithm to determine if there exists an integer i such that $A_i = i$ in an array of integers $A_1 < A_2 < A_3 < \cdots < A_N$. What is the running time of your algorithm?

2.16 Write an alternative `gcd` algorithm based on the following observations (arrange so that $a > b$):

a. $gcd(a, b) = 2gcd(a/2, b/2)$ if a and b are both even.
b. $gcd(a, b) = gcd(a/2, b)$ if a is even and b is odd.
c. $gcd(a, b) = gcd(a, b/2)$ if a is odd and b is even.
d. $gcd(a, b) = gcd((a + b)/2, (a - b)/2)$ if a and b are both odd.

2.17 Give efficient algorithms (along with running time analyses) to

a. Find the minimum subsequence sum.
* b. Find the minimum *positive* subsequence sum.
* c. Find the maximum subsequence *product*.

2.18 An important problem in numerical analysis is to find a solution to the equation $f(X) = 0$ for some arbitrary f. If the function is continuous and has two points *low* and *high* such that $f(low)$ and $f(high)$ have opposite signs, then a root must exist between *low* and *high* and can be found by a binary search. Write a function that takes as parameters f, *low,* and *high* and solves for a zero. What must you do to ensure termination?

2.19 The maximum contiguous subsequence sum algorithms in the text do not give any indication of the actual sequence. Modify them so that they return in a single object the value of the maximum subsequence and the indices of the actual sequence.

2.20 a. Write a program to determine if a positive integer, N, is prime.
b. In terms of N, what is the worst-case running time of your program? (You should be able to do this in $O(\sqrt{N})$.)
c. Let B equal the number of bits in the binary representation of N. What is the value of B?
d. In terms of B, what is the worst-case running time of your program?
e. Compare the running times to determine if a 20-bit number and a 40-bit number are prime.
f. Is it more reasonable to give the running time in terms of N or B? Why?

* **2.21** The Sieve of Eratosthenes is a method used to compute all primes less than N. We begin by making a table of integers 2 to N. We find the smallest integer, i, that is not crossed out, print i, and cross out i, $2i$, $3i, \ldots$. When $i > \sqrt{N}$, the algorithm terminates. What is the running time of this algorithm?

2.22 Show that X^{62} can be computed with only eight multiplications.

2.23 Write the fast exponentiation routine without recursion.

2.24 Give a precise count on the number of multiplications used by the fast exponentiation routine. (*Hint:* Consider the binary representation of N.)

2.25 Programs A and B are analyzed and found to have worst-case running times no greater than $150N \log_2 N$ and N^2, respectively. Answer the following questions, if possible:

a. Which program has the better guarantee on the running time for large values of N ($N > 10{,}000$)?
b. Which program has the better guarantee on the running time for small values of N ($N < 100$)?
c. Which program will run faster *on average* for $N = 1{,}000$?
d. Is it possible that program B will run faster than program A on *all* possible inputs?

2.26 A majority element in an array, A, of size N is an element that appears more than $N/2$ times (thus, there is at most one). For example, the array

$$3, 3, 4, 2, 4, 4, 2, 4, 4$$

has a majority element (4), whereas the array

$$3, 3, 4, 2, 4, 4, 2, 4$$

does not. If there is no majority element, your program should indicate this. Here is a sketch of an algorithm to solve the problem:

First, a candidate *majority element is found (this is the harder part). This candidate is the only element that could possibly be the majority element. The second step determines if this candidate is actually the majority. This is just a sequential search through the array. To find a candidate in the array, A, form a second array, B. Then compare A_1 and A_2. If they are equal, add one of these to B; otherwise do nothing. Then compare A_3 and A_4. Again if they are equal, add one of these to B; otherwise do nothing. Continue in this fashion until the entire array is read. Then recursively find a candidate for B; this is the candidate for A (why?).*

a. How does the recursion terminate?
⋆b. How is the case where N is odd handled?
⋆c. What is the running time of the algorithm?
d. How can we avoid using an extra array, B?
⋆e. Write a program to compute the majority element.

2.27 The input is an N by N matrix of numbers that is already in memory. Each individual row is increasing from left to right. Each individual column is increasing from top to bottom. Give an $O(N)$ worst-case algorithm that decides if a number X is in the matrix.

2.28 Design efficient algorithms that take an array of positive numbers `a`, and determine:
a. the maximum value of `a[j]+a[i]`, with `j` $\geq$ `i`.
b. the maximum value of `a[j]-a[i]`, with `j` $\geq$ `i`.
c. the maximum value of `a[j]*a[i]`, with `j` $\geq$ `i`.
d. the maximum value of `a[j]/a[i]`, with `j` $\geq$ `i`.

⋆**2.29** Why is it important to assume that integers in our computer model have a fixed size?

2.30 Consider the word puzzle problem on page 2. Suppose we fix the size of the longest word to be 10 characters.

a. In terms of *R* and *C*, which are the number of rows and columns in the puzzle, and *W*, which is the number of words, what are the running times of the algorithms described in Chapter 1?

b. Suppose the word list is presorted. Show how to use binary search to obtain an algorithm with significantly better running time.

2.31 Suppose that line 15 in the binary search routine had the statement `low = mid` instead of `low = mid + 1`. Would the routine still work?

2.32 Implement the binary search so that only one two-way comparison is performed in each iteration.

2.33 Suppose that lines 15 and 16 in algorithm 3 (Fig. 2.7) are replaced by

```
15        int maxLeftSum  = maxSumRec( a, left, center - 1 );
16        int maxRightSum = maxSumRec( a, center, right );
```

Would the routine still work?

***2.34** The inner loop of the cubic maximum subsequence sum algorithm performs $N(N+1)(N+2)/6$ iterations of the innermost code. The quadratic version performs $N(N + 1)/2$ iterations. The linear version performs N iterations. What pattern is evident? Can you give a combinatoric explanation of this phenomenon?

References

Analysis of the running time of algorithms was first made popular by Knuth in the three-part series [5], [6], and [7]. Analysis of the *gcd* algorithm appears in [6]. Another early text on the subject is [1].

Big-Oh, big-omega, big-theta, and little-oh notation were advocated by Knuth in [8]. There is still no uniform agreement on the matter, especially when it comes to using $\Theta()$. Many people prefer to use $O()$, even though it is less expressive. Additionally, $O()$ is still used in some corners to express a lower bound, when $\Omega()$ is called for.

The maximum subsequence sum problem is from [3]. The series of books [2], [3], and [4] show how to optimize programs for speed.

1. A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms,* Addison-Wesley, Reading, Mass., 1974.
2. J. L. Bentley, *Writing Efficient Programs,* Prentice Hall, Englewood Cliffs, N.J., 1982.
3. J. L. Bentley, *Programming Pearls,* Addison-Wesley, Reading, Mass., 1986.
4. J. L. Bentley, *More Programming Pearls,* Addison-Wesley, Reading, Mass., 1988.
5. D. E. Knuth, *The Art of Computer Programming, Vol 1: Fundamental Algorithms,* 3d ed., Addison-Wesley, Reading, Mass., 1997.
6. D. E. Knuth, *The Art of Computer Programming, Vol 2: Seminumerical Algorithms,* 3d ed., Addison-Wesley, Reading, Mass., 1998.
7. D. E. Knuth, *The Art of Computer Programming, Vol 3: Sorting and Searching,* 2d ed., Addison-Wesley, Reading, Mass., 1998.
8. D. E. Knuth, "Big Omicron and Big Omega and Big Theta," *ACM SIGACT News,* 8 (1976), 18–23.