"On two occasions I have been asked, 'If you put into the machine wrong figures, will the right answers come out?' I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question."

– Charles Babbage (1791-1871)
Inventor of the first mechanical computer

"In theory, theory and practice are the same. In practice, they are not."

– Yogi Berra

Lecture I
OUTLINE OF ALGORITHMICS

This first chapter is mostly informal. The rest of this book has no dependency on this chapter, save for the definitions in §7 concerning asymptotic notations. Hence a light reading may be sufficient. We recommend re-reading this chapter after finishing the rest of the book, when many of the remarks here may take on more concrete meaning.

There are two Appendices: Appendix A collects basic mathematical concepts that are used throughout the book. Appendix B gives a formal definition of the RAM Model of Computation.

In computer science, we study problems that can be solved on computers. Such problems can be roughly classified into problems-in-the-large and problems-in-the-small. The former is associated with large software systems such as an airline reservation system, compilers or text editors. The latter[1] is identified with mathematically well-defined problems such as sorting, multiplying two matrices or solving a linear program. The methodology for studying such "large" and "small" problems are quite distinct: Algorithmics is the study of the small problems and their algorithmic solution. In this introductory lecture, we present an outline of this enterprise.

Algorithmics is about "small" problems

Throughout this book, **computational problems** (or simply "problems") refer to problems-in-the-small. It is the only kind of problem we address. We assume the student is familiar with computer programming and has a course in data structures and some background in discrete mathematics.

¶1. Book Organization. The chapters in this book are numbered by capital roman numerals: I, II, III, IV, etc. The chapters are organized into sections, denoted §1, §2, §3, etc. Occasionally, we have subsections such as §3.1, §3.2, etc. Independent of the sections and subsections, there is a parallel organization of each chapter into *numbered paragraphs* that are numbered sequentially, starting with ¶1, ¶2, ¶3, etc. Optional or advanced sections or paragraphs are marked by an asterisk, as in §*2 or ¶*37. When referring to sections or paragraphs

[1] If problems-in-the-large is macro-economics, then the problems-in-the-small is micro-economics.

across in another chapter, we will write, e.g., §*III*.8 (section 8 in Chapter III) or ¶*IX*.3 (paragraph 3 in Chapter IX). Note that we use[2] colored fonts. Our pdf file also contains hyperlinks.

§1. What is Algorithmics?

Algorithmics is the systematic study of efficient algorithms for computational problems; it includes techniques of algorithm design, data structures, and mathematical tools for analyzing algorithms.

Why is algorithmics important? Because algorithms is at the core of all applications of computers. These algorithms are the "computational engines" that drive larger software systems. Hence it is important to learn how to construct algorithms and to analyze them. Although algorithmics provide the building blocks for large application systems, the construction of such systems usually require additional non-algorithmic techniques (e.g., database theory) which are outside our scope.

¶2. We can classify algorithmics according to its applications in subfields of the sciences and mathematics: thus we have computational *geometry*, computational *number theory*, computer *algebra*, computational *finance*, computational *physics*, and computational *biology*, etc. More generally, we have "computational X" where X can be any discipline. But another way to classify algorithmics is to look at the generic tools and techniques that are largely independent any discipline. Thus, we have sorting techniques, graph searching, string algorithms, dynamic programming, numerical techniques, etc, that cut across individual disciplines. Thus we have identified two dimensions along which the field of algorithmics can be classified. We represent these two orthogonal classification schemes using this matrix:

	geometry	topology	finance	physics	biology	algebra	...
sorting	✓	✓	✓	✓	✓	✓	
graph searching	✓		✓		✓		
string algorithms		✓		✓	✓		
dynamic programming	✓				✓	✓	
numerical methods			✓	✓			
⋮	⋮						

Computer Science is row-oriented

So each computational X is represented by a column in this matrix, and each computational technique is represented by a row. Each check mark indicates that a particular computational technique is used in a particular discipline X. Individual scientific disciplines take a column-oriented view, but *Computer Science (and also this book) takes the row-oriented view of this matrix*. In other words, we study the common elements that characterize each row. These row labels can be grouped into four basic themes:

(a) data-structures (e.g, linked lists, stacks, search trees)

[2] Students who have trouble seeing colors may request a no-color version.

(b) algorithmic techniques (e.g., divide-and-conquer, dynamic programming)

(c) basic computational problems (e.g., sorting, graph-search, point location)

(d) analysis techniques (e.g., recurrences, amortization, randomized analysis)

These themes interplay with each other. For instance, some data-structures naturally suggest certain algorithmic techniques (e.g., graphs requires graph-search techniques). Or, an algorithmic technique may entail certain analysis methods (e.g., divide-and-conquer algorithms require recurrence solving). The field of complexity theory in computer science provides some unifying concepts for algorithmics; but complexity theory is too abstract to capture many finer distinctions we wish to make. Thus algorithmics often makes domain-dependent assumptions. For example, in the subfield of computer algebra, the complexity model takes each algebraic operation as a primitive, while in the subfield of computational number theory, these algebraic operations are reduced to some bit-complexity model primitives. In this sense, algorithmics is more like combinatorics (which is eclectic) than group theory (which has a unified framework). Students may initially find this eclectic nature of algorithmics confusing. But ultimately, we hope the student will develop an "algorithmic frame of mind" that sees an over-arching unity in this jumble of topics.

§2. Teaching Algorithmics: Interlude

Although this Chapter is "introductory", it is quite abstract in its conception. The abstraction comes from our trying to connect algorithms to broader topics about general computation and the foundational questions about complexity. It is easy to become feel unmoored by the abstractions. This interlude suggests some antidote to this feeling.

Algorithms is at the heart of this book. But what is an "algorithm"? Students of this book have some prior understanding of algorithms already. They should be familiar with a programming language like `Java`, and probably an introductory course urges the student to equate "algorithm" with "program". In §4 below we will in effect say "no, that equation is not quite right" (see ¶11). But in this section we want to say "yes, that is a good start". A program is a concrete thing but an algorithm is an abstraction. It will take the next two sections of this chapter to say how they are related. In effect we tell the student – yes, your understanding of algorithm is "buggy", but view the buggy understanding in a positive light, as a necessary start. Teaching the child "one apple" and "one car" is the road to teaching the concept "one". Yet ultimately, no mathematician or philosopher[3] can really define "one" except within rather circumscribed settings.

The main goal of this book is to teach you a variety of specific algorithms. For each algorithm, we want to build your intuitions about the algorithm. One way to do this is to ask you to implement the algorithm by writing a program in a concrete programming language like `Java`. We definitely do this in our undergraduate classes. Here I want to add a caveat about programming in an Object-Oriented Programming Language (OOPL) like `Java`: *you may have to drop many common advice about information hiding practices.* For instance, in your `Java` class, you are probably told to hide the members (i.e., variables) in the class by making them private or protected, and to write setters-and-getters for each member. These protection concerns are only relevant if you are writing programs for a library for use by a general public. For teaching algorithms, this is all wrong. Forget about using protection and

[3] The Father of Modern Logic, Gottlob Frege (1848-1925) began his famous book *Grundlagen* by asking what is the concept of "one". At the end of his book that founded modern predicate calculus, he has not really answered the question either.

information hiding. They bloat your code and obscure the underlying algorithmic ideas. In our courses, students are asked to write programs for understanding, by making the logic as simple and minimalist as possible.

Unfortunately, for practical reasons, we usually drop programming in a graduate level course on algorithms. But we do have a recourse – we ask you to implement the algorithms in an informal programming language that we call **pseudo-code** (see Appendix ¶A.11 and ¶A.12). Here we stress that the pseudo-code is *not* viewed as a poor substitute for actual programming code. It is positively viewed. Why? The short answer is that pseudo-code connects more directly with our human intuition, and avoids mechanical operations which are necessary for machines, but quite unfit for humans. We are not automatons.

Instead of programming, another tool for building algorithmic intuitions is **hand simulations**. Suppose we ask you to demonstrate your understanding of the sorting problem by answering this question: *what is the sort of the array of numbers* $(3, 1, 4, 1, 5, 9)$*?* You produce the correct answer: $(1, 1, 3, 4, 5, 9)$. How did you do this? Probably by **mental simulation** of some unknown process that we might call "pre-algorithm". I have no idea what your pre-algorithm is, but if you produce the correct answer, I have some confidence that you are on the right track. If the array were somewhat bigger (say with 30 digits of π), you might have to resort to some paper-and-pencil work which would qualify as "hand-simulation" of your pre-algorithm.

But proper hand simulations begins when we have some specific algorithm such as **Merge Sort** to test. If we ask you to do hand simulation of Merge Sort on $(3, 1, 4, 1, 5, 9)$, you need to produce various intermediate products on the way to the final sorted output. In this case we must agree on some format for representing your hand-simulation. In the case of Merge Sort, we probably ask you to organize the intermediate sorts of subarrays in the form of a recursion tree: recursively sort $(3, 1, 4)$ and $(1, 5, 9)$ to produce $(1, 3, 4)$ and $(1, 5, 9)$, then merge them into $(1, 1, 3, 4, 5, 9)$.

Like hand simulations, another handy skill is **pen-paper estimates** that the physicists are famous for. Here, students should freely use the "Computer Science estimates" such as

$$2^{10} \simeq 1000 = 1K, \quad 2^{20} \simeq 10^6 = 1M.$$

Given any big quantity T 1, we should be able to state its **order of magnitude** of T (which we define as the integer closest to $\log_{10} T$). E.g., the order of magnitude of $T = 2^{20}$ is 6.

Summary: it is human intuitions about algorithms that we seek to inculcate. This is like building your "algorithmic muscles" so that you can go on to tackle new and unknown algorithmic problems. The formal tools, data structures and actual algorithms must be built on this foundation. A more specific goal is to build insights through mathematical analysis. Example of such activities include proving correctness of an algorithm and analysis of its complexity. Although every student should try to get a taste of this activity, some of this may be more appropriate for the mathematically-inclined or PhD students. We ask the student's indulgence and to simply pass over such parts (usually the asterisked sections or paragraphs).

§3. Computational Problems: What are we solving?

Despite its name, the starting point for algorithmics is not algorithms, but **computational problems**. But what are "computational problems"? We mention three main categories.

¶3. (A) Input-output problems. Such problems are the simplest to understand. A **computational problem** is a precise specification of input-and-output (I/O) formats, and for each input instance I, we have a set $\mathcal{O}(I)$ of possible output instances. The word "format" emphasizes the fact the representation of the input I and outputs $\mathcal{O}(I)$ are part and parcel of the problem. Inputs that are improperly formatted are meaningless. In practice, standard formatting may be taken for granted (e.g., numbers are assumed to be in binary or decimal notation, and set elements are arbitrarily listed without repetition). Note that the input-output relationship need not be functional: a given input may have several acceptable outputs when $|\mathcal{O}(I)| > 1$. Here are two examples of such problems: **(A1) Sorting Problem.** The input is a sequence of numbers $(a_1, \ldots, a_n)$ and output is a rearrangement of these numbers in[4] increasing order, $(a'_1 \leqslant \cdots \leqslant a'_n)$. An input instance is $(2, 5, 2, 1, 7)$, with corresponding output instance $(1 \leqslant 2 \leqslant 2 \leqslant 5 \leqslant 7)$.

The epigraph in front of this chapter by Charles Babbage talks about meaningless inputs.

(A2) Primality Testing. Input is a natural number n and output is either YES (if n is prime) or NO (if n is composite). Numbers are assumed to be encoded in decimal. E.g., if the input is 123 then the output is NO. But for the input 23, the output is YES. This is an example of a **decision or recognition problem**, where the output have only two possible answers (YES/NO, 0/1, Accept/Reject).

Simplest imaginable type of problem?

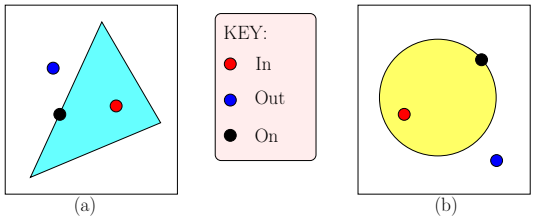


Figure 1: Classifying points relative to (a) Triangles, (b) Discs

A generalization of decision problems is to have outputs that comes from a fixed finite set. For instance, in computational geometry, the decision problems tend to have three possible answers: Positive/Negative/Zero or IN/OUT/ON. For instance, the **point classification problem** is where we are given a point and some geometric object such as a triangle or a cell. The point is either inside the cell, outside the cell or on the boundary of the cell. See illustration in Figure 1.

¶* 4. (B) Preprocessing problems. In such problems, *given a set X of objects, we must construct a data structure $D(X)$ such that, for any **query** (of a suitable type) about X, we can use $D(X)$ to efficiently answer the query.* There are two stages in such problems: a preprocessing stage and a query stage. In the query stage, we are asked to answer some unknown number k of queries. Typically we do not know k, and we often think of k as very large. We must provide an algorithm for each stage: a preprocessing algorithm and a querying algorithm. Here are two examples:

Two-staged problems

(B1) Ranking Problem. The preprocessing input is a set X of numbers. A query on X is a number q for which we like to determine its rank in X, denoted $rank(q, X)$. We define $rank(q, X)$ to be the number of items in X that are smaller than or equal to q. E.g., if $X = \{3, 2, 7, 5\}$ and $q = 6$ then $rank(q, X) = 3$. If $|X| = n$, the standard solution is for $D(X)$ to

[4] In this book, **increasing order** really means "non-decreasing order with ties broken arbitrarily". A list $(a_1, \ldots, a_n)$ is increasing order is usually denote $(a_1 \leqslant \cdots \leqslant a_n)$. We say **strictly increasing order** in case we disallow ties in the list. This is then denoted $(a_1 < \cdots < a_n)$. A similar convention applies to **decreasing order**.

be a sorted array $A[1..n]$ containing the elements of X in increasing order. To answer the query $rank(q, X)$, we perform a binary search to determine the unique i such that $A[i] \leqslant q < A[i+1]$ (with $A[0] = -\infty$ and $A[n+1] = \infty$). Thus the preprocessing amounts to sorting the set X, and querying is solved with a binary search algorithm.

(B2) Post Office Problem. Many problems in computational geometry and database search are the preprocessing type. The following is a geometric-database illustration: given a set X of points in the plane, find a data structure $D(X)$ such that for any query point q, we find an element in X that is closest to q. Think of X as a set of post offices and we want to know the nearest post office to any position q. E.g., X are cities in The Netherlands as shown in Figure 2. The datastructure $D(X)$ is some representation of the Voronoi diagram of X which subdivide the plane into polygonal regions (or "zones"), each region comprising the points closest to a city. The best solution for this Post Office Problem can be viewed as a 2-dimensional generalization of our solution to the ranking problem.

Figure 2: Post-Office zones in The Netherlands (see [3, pp. 147–163])

The complexity of a solution to the preprocessing problem can be described by a triple of complexity functions: $(S(n), P(n), Q(n))$ where $S(n)$ is the **space complexity** of the datastructure $D(S)$ when $|S| = n$, $P(n)$ is the **preprocessing time** to construct $D(S)$, and $Q(n)$ is the worst case **query time**. Thus, the above solution for the ranking problem has the complexity triplet

$$(S(n), P(n), Q(n)) = O(n, n \log n, \log n). \tag{1}$$

Here, we write "$O(f, g, h)$" instead of $(O(f), O(g), O(h))$ for complexity triplets. It turns out that the same triplet can be achieved for the Post Office Problem! This may sound surprising since the Post Office Problem seems much harder. One can imagine two solutions that exhibit a tradeoff among the 3 complexities functions (see Exercises).

¶* 5. (C) Dynamization and Online problems. Now assume the input S is a set of objects. For example, a database might be regarded as such a set. In addition to queries on S, we now allow operations that add or remove objects from S. Such operations are called **updates**. We will still have **queries** on S as in preprocessing problems. Collectively, updates and queries are called **requests**. In a **dynamization problem**, we are given a sequence

$$(r_1, r_2, \ldots, r_m)$$

of requests. Initially, $S = \varnothing$ is the empty set, and we construct the initial data structure $D(\varnothing)$. Subsequently, we respond to each request r_i in order: $i = 1, 2, \ldots, m$. Let S_i be the set of objects in S after the ith request. If r_i is an update to add (resp., to remove) an object x, then $S_{i+1} = S_i \cup \{x\}$ (resp., $S_i \backslash \{x\}$). Otherwise, r_i is a query and $S_{i+1} = S_i$. Each request must be answered in the given order (this is called the **online** requirement). In particular, we are not allowed to answer the requests in batches. Let S denote a **dynamic set** whose membership can change over time. The **state** of S at any time i is just the set of objects in S

at time i. Relative to the sequence $(r_1, \ldots, r_m)$, let S_i be the state of S at time i (i.e., after answering request r_i). Similarly, the data structure $D(S)$ can now change over time: let D_i denote the state of $D(S)$ at time i. Here are three examples:

(C0) Set Maintenance Problem. The updates are either to add or remove an object from S. Queries have the form "is x in S?" for any object x. Our above preprocessing problem for a (static) set S with queries $q_1, \ldots, q_m$ can be viewed as a special case of the set maintenance problem: if $S = \{x_1, \ldots, x_n\}$, then the preprocessing problem amounts to processing the requests sequence

$$(r_1, \ldots, r_n, q_1, \ldots, q_m)$$

where r_i amounts to adding x_i.

(C1) Dynamic Ranking Problem. This is a generalization of the ranking problem (B1) in which the queries are just rank queries. In other words, we have updates on S intermixed with rank queries on S. In the static problem (B1) above, the data structure $D(S)$ is a sorted array. This will not be very efficient under updates of S. The preferred data structure here is the **dynamic search trees** (see Chapter III).

(C2) Graph Maintenance Problems. Dynamization on graphs can be viewed as a set maintenance problem in which we are maintaining two sets: a set V of vertices and a set E of edges where (V, E) represents a graph. The graph can be directed or undirected. The queries can be any kind of query on graphs. For instance, in the **dynamic connected component problem**, the queries are defined by a pair $u, v \in V$ of vertices. The answer is either YES or NO, depending on whether u, v are in the same connected component of (V, E) or not.

¶6. (D) Promise Problems. When introducing the computational problems above, we stressed that each input I must be properly "formatted", i.e., follows a pre-specified form. This is a syntactical requirement and is a *sine qua non*. But very often we encounter problems where the interpretation of I as a mathematical object has additional requirements that cannot be excluded by the syntax. E.g., if I represents a polynomial $f(x)$, we might require[5] that $f(x)$ has only "simple" roots (mathematically, it means that if $f(\alpha) = 0$ then $f'(\alpha) \neq 0$). This requirement cannot be checked syntactically. When the correctness of an algorithm depends on such semantic promises about its input, we call it a **promise algorithm**; the corresponding problem it solves is called a **promise problem**. In practice, promise algorithms are very useful because they can be deployed in situations where we know that the promise can be kept. Very often, there are algorithms that can check whether I satisfies a promise. E.g., $f(x)$ has only simple roots iff $\texttt{GCD}(f, f') = 1$. So, check this promise requires algorithms to compute the derivative f' and to compute GCD of polynomials. But very often, it is not at all clear that there are algorithms to check the promise (this happens in numerical algorithms for differential equations). In this book, we will encounter promise algorithms (e.g., in graph algorithms, we might assume that the input graph is connected.)

¶7. (E) Heuristics and Pseudo-problems. Let us illustrate what we regard to be a "pseudo-problem" from the viewpoint of our subject. Suppose your boss asks your IT department to "build an integrated accounting system-cum-employee database". This may be a real world scenario but it is not a legitimate topic for algorithmics because part of the task is to figure out what the input and output of the system should be, and there are probably other implicit non-quantifiable criteria (such as available technology and economic realities). This

[5] There are root finding algorithms that depends on this particular promise.

is also a good place to inject a warning: it is popular today to refer to almost any computer program as an "algorithm" (e.g., "Google's algorithm"). But unless we have some objective criteria to say whether the output of the program is correct or not, it is not an algorithm in the sense of this book. We prefer to call them **heuristics**. Most "algorithms" of Artificial Intelligence, Machine Learning or Data Science are heuristics. Heuristics are very useful tools, especially when they are tuneable, with a human in the loop. We cannot leave all decisions to the heuristics when the task is *critical* (like a mission to Mars). Fortunately most applications of heuristic are *noncritical* (e.g., guessing what movies you like based on your clicks).

Example: suppose I want an image of a real tree to illustrate my lecture on Binary Search Trees. So I search Google's database of images for "trees". In the pages of search results, I will inevitably detect images that should never regarded as "tree image". These are are the **false positives** which are visible for us to see. But what about the **false negatives**, i.e., images of bona-fide trees in the database that are not shown? (Perhaps I was hoping to get just one of these images.) In any case, it is hard to characterize the exact set of images that ought to be returned: Is a forest image OK? Partially occluded trees? How much occlusion is acceptable? Trees mostly in shadows? Just leaves? Bark? In short, this problem has no precise notion of correct output. On the other hand, it is also not a critical application.

Exercises

Exercise 3.1: Please give algorithms in pseudo-code for the problems (A1), (A2), (B1), (B2), (C1), (C2) in this section. Do not worry about efficiency. Nevertheless, give the time complexity of your solutions. NOTE: for preprocessing problems (B1,B2), you need to give the complexity of the pre-processing algorithm and well as the query algorithm. ♢

Exercise 3.2: We have described a solution to the ranking problem (B1) (see §I.4) whose triplet complexity is

$$O(n, n \log n, \log n). \tag{2}$$

[eqn.(1) in the text]
(a) Give a different solution to this problem that exhibits a trade-off relative to (2). State the complexity triplet of your solution.
(b) Are there scenarios where your solution in (a) better preferable over (2)? HINT: suppose there are k queries in the Query Stage. What if k is relatively small, e.g., $k = O(\log n)$?

♢

End Exercises

§4. Computational Model: How do we solve problems?

¶8. Suppose we want to solve a computational problem P where P falls under one of the 3 problem classes (A), (B) or (C) in the previous section. Now we must choose the tools for solving it. This is given by the **computational model**. Any conventional programming language such as `C++` or `Java` (suitably abstracted) can be regarded as a computational model. A computational model is specified by

(a) the kind of data objects that it deals with

(b) the primitive operations to operate on these objects

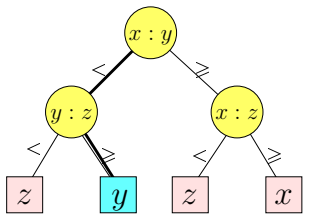
(c) rules for composing primitive operations into larger units called **programs**.

Programs can be viewed as individual instances of a computational model. For instance, the Turing model of computation is an important model in complexity theory and will be introduced in the last Chapter of this book. The programs in this model are called Turing machines. The Turing model "universal" in the sense that any solvable problem can be solved by some Turing machine. In Appendix B, we describe another universal model called the Random Access Model (RAM). But in this section, we want to introduce 3 simple computational models. These are clearly not universal, but we can use them to illustrate some general concepts in complexity.

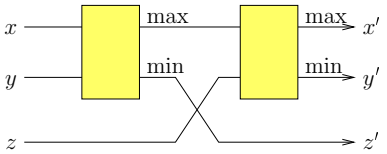
¶9. Models for Comparison-based Problems. The sorting problem has been extensively studied since the beginning of Computer Science (from the 1950's). Sorting is just a representative of a whole class of problems that can be solved using the primitive capability of comparing two elements. It turns out that there are several distinct computational models for such problems. We will next describe three of them: the **comparison tree model**, the **comparator circuit model**, and the **tape model**. In each model, the data objects are **elements** from a linear order. E.g., the elements may be integers from $\mathbb{Z}$ or real numbers from $\mathbb{R}$ or ASCII strings with the usual ASCII sorting order.

3 sorting models

¶10. Comparison Tree Model. First we consider the comparison tree model which has only one primitive operation, viz., comparing the two elements x, y resulting in one of two outcomes $x < y$ or $x \geqslant y$. Such a comparison is usually denoted "$x : y$". We compose these primitive comparisons into a **tree program** by putting a comparison at the internal node of a binary tree. Tree programs only represent flow of control and are, more generally, are called **decision trees** where the decision at each internal node is based on a predicate (such as a comparison). Figure 3(a) illustrates a comparison tree on inputs x, y, z.



(a) Comparison Tree



(b) Comparator Circuit

Figure 3: Two programs to find the maximum of x, y, z.

To use a comparison tree, we begin at the root, and perform the indicated comparison, say $x : y$. If $x < y$, we proceed to the left child; otherwise, we proceed to the right child. We continue recursively in this manner until we reach a leaf, and stop. Let us illustrate this with the comparison tree in Figure 3(a) (follow the thick path from root to a leaf). Suppose our input is $(x, y, z) = (7, 9, 3)$. Then the comparison $x : y$ at the root tells us to compare $7 : 9$. Since $7 < 9$, we move to the left child. The comparison at the left child is $y : z$, i.e., $9 : 3$. Since $9 \geqslant 3$, we move to the right child. We have reached a leaf. This leaf specifies the element y

(i.e., 9) which is our output. The reason for this output is that our comparison tree is supposed to be an algorithm to find a maximum of x, y, z.

So the outputs of a comparison tree are specified at each leaf. For instance, if the tree is meant for sorting, each leaf will output the sorted order of the input set. These examples raise the question: what exactly is the nature of the output at each leaf? There is a precise answer: *the output at each leaf must be determined from the set of relations collected along the edges of the path to the leaf.* Let us unroll this remark: each edge of the comparison tree represents a relationship of the form $x < y$ or $x \geqslant y$, and the set of all these relationships along a path to any node v forms a partial order $\phi(v)$ on the input set. *The tree program solves problem P if for each leaf v, the partial order $\phi(v)$ determines some solution to P.* Adopting a notation from mathematical logic, we may write

$$\phi(v) \models P \tag{3}$$

to indicate this relationship. In hand simulations, it is helpful draw the Hasse diagram (Appendix §A.4) of $\phi(v)$ next to each node v.

Example: For the comparison tree Figure 3(a), check that the path taken by the input $\{x = 7, y = 9, z = 3\}$ ends at a leaf v where $\phi(v) = \{x < y, y \geqslant z\}$. Clearly, $\phi(v)$ determine y (the output) as a maximum. The outputs at the other 3 leaves of the tree can be similarly verified.

¶11. Comparison Circuit Model. We come to the second computational model for sorting-like problems: we again have only one kind of primitive called a **comparator** that takes two input elements x, y and returns two outputs, $\max\{x, y\}$ and $\min\{x, y\}$. Each comparator can be viewed as a node in a graph with two incoming edges and two outgoing edges. Graphical conventions: *inputs enter from the left and exit to the right of each comparator; the maximum output is drawn above the minimum output.* Comparators are composed into **(comparator) circuits** as illustrated in Figure 3(b). This circuit has inputs x, y, z and outputs x', y', z' and two comparators (shown as rectangles). For instance, the output x' in the circuit Figure 3(b) is the maximum of x, y, z; if the circuit is regarded as an algorithm for computing the maxima, then we can ignore the outputs y', z'. For the sorting problem, no output may be ignored.

Formally, a **comparator graph** is a directed acyclic graph (DAG) whose nodes are classified as **input**, **output** or **comparator**. Each type of node is characterized by their (in,out)-degrees: $(0, 1)$ for input, $(1, 0)$ for output, and $(2, 2)$ for comparators. The output from each comparator may be connected to the input of another comparator by a **wire**. When an input (resp. output) has no incoming wire (resp. outgoing wire), it is considered an input (resp., output) of the entire circuit. It follows that the number of input and output nodes of the circuit are equal: denote them by $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$. Note: we may sometimes ignore some outputs. E.g., to compute the maximum of $x_1, \ldots, x_n$, we are only interested in one output, say y_1. Note that a comparator graph has captured all the essentials of a comparator circuit except one: to make the graph equivalent to a circuit, we need to "order" the two output wires from each comparator node: we must designate one of the output wires as "max" and the other as "min".

What is the semantics? The inputs are to be instantiated with elements from any totally ordered set (say, real numbers). We can label each wire with a min-max expression over the input variables $x_1, \ldots, x_n$. If the two input wires to a comparator has the expression E_1, E_2 then the output wires are labeled $\min(E, E')$ and $\max(E, E')$, consistent with the above min/max output designations. For instance, in Figure 3(b), the labels x', y', z' can be interpreted as these

min-max expressions:

$$x' = \max(\max(x, y), z), y' = \min(\max(x, y), z), z' = \min(x, y).$$

It is clear that the comparator circuit compute permutations of its input: $(y_1, \ldots, y_n)$ is a permutation of $(x_1, \ldots, x_n)$.

¶12. Tape Model. The third model for sorting-like problems is the Tape Model, studied in [5]. Each program in this model uses some finite number $k \geqslant 1$ of **tapes**. Typically $k \leqslant 6$. Each tape has a finite sequence of **slots**, indexed sequentially from $0, 1, \ldots, n$. The value n may not be known and may be different for each tape. Each slot can hold an **item** or is **empty**. Each tape has a **head** that is at some position $i \in [0..n]$ at any moment: we can test whether the ith slot is empty, and if non-empty, we can read the item in the ith slot into some variable x or we can write an item in x into this slot. We call the read/write operation an **advance operation** because immediately after the operation, the head moves to position $i+1$. There is a **reset operation** that puts the tape head in the initial position, $i = 0$. Finally, we can also test if we are at the end of the tape (is $i = n$?).

Below, we will use such primitives to construct an algorithm to merge the contents of two input tapes.

> One model of how tapes work is the audio cassette tape (remember those?). Each end of the cassette tape is attached to a separate spool, and the two spools are positioned at a fixed distance apart. The tape is wound up on each spool so that "free tape" between the two spools is held taut: the head position is placed somewhere on the free tape. Unwinding the tape on one spool requires a corresponding winding up on the other spool to keep this free tape taut. To "reset" means to completely unwind one of the spools. Unlike a cassette tape, a computer tape is wound up on one spool only. To read/write on such a tape, we need to first physically attach the free end of the computer tape to an empty spool, and operate it like a cassette tape.

The tape model was very important in the early days of computing when main memory was expensive and physical tapes were cheap and became the standard medium for storing large data sets. The tape technology appears to be obsolete by around 2000. Interestingly, with the advent of the web-age, a variant of this model called **streaming data model** is coming back. Now we are faced with a huge volume of real-time sensing data from sources such as satellites and other monitoring devices. These data are often simply stored for future researchers. Instead of sorting, we need to compute some simple transformation of the data. The large volume of data implies that we cannot afford to make more than one or two passes over the data to do this computation. This is like the constraints of the tape model. For the class tape model, We refer to Knuth [5, Chap. 5.4], under external sorting.

¶13. From Programs to Algorithms. To recap, we started with a problem P, and chose/design a program A in some model M. For A to solve P, we must make sure there is a match between the data objects in the problem specification and the data objects handled by model M. If not, we must specify some suitable encoding of the former objects by the latter. After making such encoding conventions, we may call A an **algorithm for** P if, for each legal input of P, the program A indeed computes a correct output. Thus the term "algorithm" is a

semantic concept, signifying a program A in its relation to some problem P. We might indicate this relation using a standard symbol from logic, "$A \models P$", extending the notation $\phi(v) \models P$ above. The program A itself is a purely syntactic object, capable of more than one interpretation. E.g., the two programs in figure 3(a,b) are interpreted as algorithms to compute the maximum of x, y, z; but it is also possible to view them as algorithms for other problems (see Exercise).

The symbols $\models$ is called a double turnstile.

¶14. The Merging Problem. A subproblem that arises in sorting is the **merge problem** where we are given two sorted lists $(x_1 \leqslant x_2 \leqslant \cdots \leqslant x_m)$ and $(y_1 \leqslant y_2 \leqslant \cdots \leqslant y_n)$ and we want to produce a sorted list $(z_1 \leqslant z_2 \leqslant \cdots \leqslant z_{m+n})$ where $\{z_1, \ldots, z_{m+n}\} = \{x_1, \ldots, x_m, y_1, \ldots, y_n\}$.

The algorithmic idea for merging is as follows: what should the first output element be? Well, it is the minimum of x_1 and y_1, decided by one comparison. Assume this output is x_1. What is the next one? Well, it must be either x_2 or y_1, and another comparison will decide. So the general picture is that, for some $i, j \geqslant 1$, we have already output $x_1, \ldots, x_{i-1}$, and we have output $y_1, \ldots, y_{j-1}$. The next output element is either x_i or y_j, as is determined by a comparison, $x_i : y_j$. This invariant is easy to maintain. When one list is exhausted, we simply output the remaining elements in the other list. Here then is our algorithm, written in a non-specific pseudo-programming language:

Merge Algorithm
Input: $(x_1 \leqslant \cdots \leqslant x_m)$ and $(y_1 \leqslant \cdots \leqslant y_n)$.
Output: The merger $(z_1 \leqslant \cdots \leqslant z_{m+n})$ of these two lists.

```
▷ Initialize:
    i ← 1, j ← 1, k ← 1.
▷ Loop:
    While (i ⩽ m ∧ j ⩽ n)
        If (x_i < y_j)
            z_k ← x_i, i++, k++.
        else
            z_k ← y_j, j++, k++.
▷ Terminate:
    If (i > m)    ◁ The x's are exhausted, output the remaining y's
        (z_k, ..., z_{m+n}) ← (y_j, ..., y_n).
    else   ◁ The y's are exhausted, output the remaining x's
        (z_k, ..., z_{m+n}) ← (x_i, ..., x_m).
```

what is pseudo-programming?

The student should note the conventions used in our programs, such as illustrated here. In nutshell, *a pseudo-program provides a clear description of the flow of control of the algorithm, without restricting how operations in non-control steps of the algorithm are described.* Typically, it means we explain these operations in English or mathematical terms. Since the flow of control (loops, branches, termination) must be explicit, the pseudo-program should specify how control variables such as Boolean flags or loop counter variables are modified. If we iterate over elements in some queue, we need to specify how the queue is initialized, modified or tested. Of course, computers are not smart enough to compile our programs. That is alright because our programs are intended for human consumption, not computers. Conversely, if we write our programs rigidly to be compilable by a computer, they would be less readable by humans. We place our priority on humans.

Here are some guidelines for pseudo-programs. First, we prefer English and mathematical notations over programming notations because the former are both more compact and more flexible. Natural languages (and English in particular) are highly effective for communication. Mathematics is usually more compact but certainly more precise. In contrast, programming notations are optimized for compilers and machines. In most computer languages, assignment is denoted "$x = y$". This conflicts with the standard meaning in mathematics where "$x = y$" is a predicate. Therefore we will use "$x \leftarrow y$" for assignment. Second, we use indentation for program blocks – this reduces clutter, improves readability. Third, like scripting languages, we do not declare our variables. The text and context should tell you how to interpret these variables. Finally, we use two kinds of comments: ($\triangleright$ *forward comments*) to describe what is coming up next, and ($\triangleleft$ *backward comments*) to briefly explain the immediately preceding code (usually on the same line).

$x \leftarrow 1$, not $x = 1$.
$x \leqslant y$, not $x <= y$.
etc.

¶* 15. Uniform versus Non-uniform Models. The preceding merge algorithm should look more familiar to students than our comparison trees. Formally, we can regard this program as belonging to the RAM (Random Access Machine) model. It is described in Appendix B, but for now the student may just identify a RAM program with program in any conventional programming language like `Java` or `C++`. Besides the familiarity factor, there is fundamental difference between the RAM model and the comparison tree model: the former is a **uniform model** and the latter is a **non-uniform model**.

Before we explain this uniform/non-uniform distinction, let us see how we can extract from the merge algorithm of ¶14 an infinite set

$$T = \{T_{m,n} : m, n \in \mathbb{N}\} \tag{4}$$

of comparison trees. Each $T_{m,n} \in T$ is a comparison tree on the sorted input sequences $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$. The root of $T_{m,n}$ has the comparison $x_1 : y_1$ because the Merge Algorithm begins with this comparison. If $x_1 < y_1$, then the Merge Algorithm will next compare $x_2 : y_1$. So we install $x_2 : y_1$ at the left child of the root. But if $x_1 \geqslant y_1$, the Merge Algorithm will next compare $x_1 : y_2$. Accordingly, we install $x_1 : y_2$ at the right child of the root. We can proceed this way to install a comparison at each of the node of an ever expanding comparison tree. But when there are no more comparisons, we have reached a leaf. We could install the sorted output $(z_1, \ldots, z_{m+n})$ at the leaf if we like (but formally, it is not necessary). Each $T_{m,n}$ is an algorithm for merging the sorted list $(x_1, \ldots, x_m)$ with the sorted list $(y_1, \ldots, y_n)$. The set T is called a **non-uniform algorithm** for merging. This process of constructing $T_{m,n}$ is known as "unrolling" the Merge Algorithm (for the indicated inputs). In the Exercise, we ask you to explicitly construct $T_{2,4}$ by this unrolling process.

unrolling a uniform algorithm

Suppose X is the input set for a problem P. Assume that we have a notion of **input size**, which is a function

$$size : X \to \mathbb{N}. \tag{5}$$

where $size(x)$ is known as the **size** of $x \in X$. *We assume there are inputs of arbitrarily large size.* E.g., for the sorting program, X is the set of all sequences of real numbers; if $x \in X$ is a sequence of n elements, then $size(x) = n$. Then we have $X = \cup_{n\in\mathbb{N}} X_n$ where $X_n := \{x \in X : size(x) = n\}$. Note that X_n might be empty for arbitrarily large values of n. For instance, if our inputs are square matrices, and we measure size of a matrix by the number of entries, it follows that X_n is empty unless n is a square (i.e., $n = m^2$ for some m).

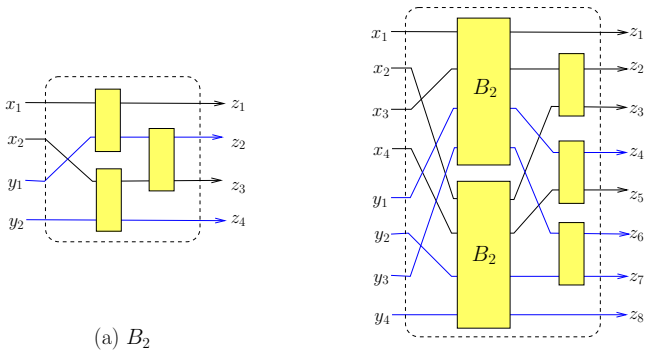
A **uniform algorithm** for P is one that accepts all $x \in X$. But algorithm A_n that accepts only inputs from X_n (for some $n \in \mathbb{N}$) is called a **finite program**. Putting together an infinite set of such finite programs,

$$A = \{A_n : n \in \mathbb{N}\}. \tag{6}$$

If each A_n solves the problem P for inputs of size n, then we call A a **non-uniform algorithm** for P. Intuitively, A is "non-uniform" because, *a priori*, there need not be any systematic method of generating the different A_n's.

For our merging problem, the input set X is now the set of all pairs (x, y) where x, y are sorted sequences. We define $size : X \to \mathbb{N}^2$ where $size(x, y) = (m, n)$ if x has length m and y length n. Clearly, the Merge Algorithm in ¶14 is a uniform algorithm for merging. The RAM model is called a "uniform model" because it permits the construction of uniform algorithms such as the Merge Algorithm. Pointer machines (see Chapter 6) and Turing machines are other examples of uniform models. In contrast, each program in the comparison tree model admits[6] a fixed size input. Thus the comparison tree model can only provide non-uniform algorithms such as (4). The relationship between complexity in uniform models and in non-uniform models is studied in complexity theory.

¶16. Merging in the Circuit Model. We now provide a comparator circuit B_n to merge two sorted lists $(x_1 \leqslant \cdots \leqslant x_n)$ and $(y_1 \leqslant \cdots \leqslant y_n)$ of size n each. It is called Batcher's Odd-Even Sort Algorithm (1968). For simplicity, assume n is a power of 2, and the construction is recursive. For the base case, B_1 is just a comparator. Figure 4 shows B_2 and B_4.



(a) B_2 (b) B_4

Figure 4: Merging Circuit

In general, B_n is comprised of two copies of $B_{n/2}$ connected to a "layer" of $n-1$ comparators as shown in the figure. The first (second) copy of $B_{n/2}$ takes as input the wires x_i and y_i where i is odd (even). For instance, the first copy of $B_{n/2}$ takes the inputs $(x_1, x_3, \ldots, x_{n-1})$ and $(y_1, y_3, \ldots, y_{n-1})$, while the second copy takes $(x_2, x_4, \ldots, x_n)$ and $(y_2, y_4, \ldots, y_n)$. Let their sorted outputs are $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ respectively. Next, let us do the following $n-1$ "compare-exchanges":

$$v_1 : u_2,\ v_2 : u_3,\ v_3 : u_4, \ldots, v_{n-1} : u_n.$$

After comparing $v_i : u_{i+1}$, we exchange (if necessary) the values of v_i and u_{i+1} so that the relation $v_i \leqslant u_{i+1}$ is true. Note that u_1 and v_n are not involved in these compare-exchanges. The final sorted output is

$$(u_1, v_1, u_2, v_2, u_3, \ldots, v_{n-1}, u_n, v_n).$$

The correctness of this circuit can be shown by the so-called "zero-one principle" (Exercise).

[6] For this purpose, we say that the "input variables" for a comparison tree is the set $\{v_1, \ldots, v_n\}$ of variables that appear in some comparison in the tree. A sequence $(x_1, \ldots, x_n)$ of numbers is regarded as "input instance" under the assignments $v_i \leftarrow x_i$ for each $i = 1, \ldots, n$.

¶* 17. Merging in Tape Model. Let us now illustrate how one can design algorithms in the tape model. We assume some conventional programming language (or RAM model), but augmented with 6 **tape primitives** on any tape T:

$$\texttt{EOT}(T), \quad \texttt{EOC}(T), \quad \texttt{READ}(T,x), \quad \texttt{WRITE}(T,x), \quad \texttt{RESET}(T), \quad \texttt{ERASE}(T) \tag{7}$$

where T is a tape and x is a item variable in a program under the Tape Model. In our application, the item may be an integer. Assume the tape is a finite array $T[0..n]$ where $T[i]$ is the ith slot. At any moment, it stores some $m \in [0..n]$ items in the first m slots, $T[0..m-1]$. The position of the tape head is restricted to $i \in [0..m]$ (i.e., it cannot go past the first empty slot which is $T[m]$). The first two primitives are predicates: $\texttt{EOT}(T)$ ("End-of-tape") returns `true` iff the head position is n, and $\texttt{EOC}(T)$ ("End-of-contents") returns `true` iff the head position is m. Thus $\texttt{EOC}(T)$ returns `false` means that the current slot holds an item. Both n (which is fixed) and m (which varies) are not directly known (of course an algorithm could determine and maintain their values if desired). The remaining primitives are operations: $\texttt{READ}(T,x)$ copies the item in the current slot on tape T into variable x, while the operation $\texttt{WRITE}(T,x)$ copies the item x into the current slot. Both `READ` and `WRITE`, when successful, advances the head to the next slot. These operations are "NO-OP" (nothing happens) if their prerequisites fail: reading an empty slot or writing to the end-of-tape. $\texttt{RESET}(T)$ rewinds the tape to the very beginning of the tape, so the head position is set to 0. Finally, $\texttt{ERASE}(T)$ simply erases the contents of the tape from the current head position i to the end-of-tape, but the head position remains at i. In effect, we set $m \leftarrow i$. Here are two examples of how to use these primitives.

E.g., to completely erase the contents of tape T, you do $\texttt{RESET}(T)$ followed by $\texttt{ERASE}(T)$.

E.g, to check if the tape T is empty, we first do $\texttt{RESET}(T)$ and then check if $\texttt{EOC}(T)$ is true.

We now provide a 3-tape algorithm to do merging: the input are placed in two tapes T_1, T_2, each containing a list of sorted items (in increasing order). We want to merge the their contents into tape T_0. It is an implementation of Merge Algorithm in ¶14 in the Tape Model. We use the variables x_i $(i = 1, 2)$ to store an item read from tape T_i. There are two Boolean variables, b_1 and b_2, where $b_i =$ `true` iff variable x_i holds an item that has not been output.

```
Tape Merge Algorithm:
▷ Initialization: set-up x_i, b_i (i = 1, 2)
    RESET(T_0),   ERASE(T_0).
    RESET(T_1),   RESET(T_2).
    b_1 ← b_2 ← true.
    If EOC(T_1) then b_1 ← false
    else READ(T_1, x_1).
    If EOC(T_2) then b_2 ← false
    else READ(T_2, x_2).
▷ Main Loop: both b_1 and b_2 are true
    While (b_1 ∧ b_2)
        If (x_1 ⩽ x_2)
            WRITE(T_0, x_1).
            If EOT(T_1) then b_1 ← false
            else READ(T_1, x_1).
        else
            WRITE(T_0, x_2).
            If EOT(T_2) then b_2 ← false
            else READ(T_2, x_2).
▷ Clean-Up: either b_1 or b_2 is false.
    While (b_1)
        WRITE(T_0, x_1).
        If EOT(T_1) then b_1 ← false
        else READ(T_1, x_1).
    While (b_2)
        ···    ◁ Repeat the previous while-loop for T_2
```

Note that this algorithm uses three tapes but in general, we can use any finite number of tapes. Our program can have any finite number of variables to store items: in this example, we use just two variables (x_1, x_2). In an Exercise, we ask you to design a tape algorithm for sorting. The goal is to minimize the number of passes (i.e., number of `RESET`'s).

¶18. Program Correctness. Recall our distinction between a "program" and an "algorithm". By definition, an algorithm is a program that is *correct* for a given problem. There is an area of computer science that formally studies program correctness, from the logical analysis of correctness, to the design of software tools for proving correctness. Correctness is also central for us. It is usual to divide the notion of "program correctness" into two parts: **partial correctness** and **halting**. The partial correctness part says that, provided the program halts, it gives the correct output. The halting part asserts that the program always halt. Halting is sometimes trivial (e.g., in our Merge algorithm above) but sometimes, it can be be highly nontrivial. We should say that there are some programs in which the "halting part" requires non-halting! For instance, if the program is an operating system, we want to ensure that it never halts. Our definition of "computational problem" precludes such non-halting algorithms.

Exercises

Exercise 4.1: We interpreted the programs in Figure 3(a) and (b) as "algorithms for finding the maximum of $\{x, y, z\}$". But the notion of an "algorithm" is a semantical concept.

So the same programs can be given different interpretations. Please give a different interpretation to these two programs. I.e., view them as solving a different problem.

NOTES: Regard the output at each leaf of a comparison tree as part of the "interpretation". So you may change the output at each leaf, but do not change the programs. However, x, y, z are still numbers – we are not interested in re-interpreting these numbers as time, strings, apples, etc. ♢

Exercise 4.2: (a) Extend the comparison tree in Figure 3(a) so that it sorts three input elements $\{x, y, z\}$.
(b) Extend the comparator circuit in Figure 3(b) so that it sorts three input elements $\{x, y, z\}$. ♢

Exercise 4.3: Let X be a finite set, and $\mathbb{P}(X)$ be the set of partial orders on X. Two partial orders $P, Q \in \mathbb{P}(X)$ are **equivalent**, denoted $P \equiv Q$ if they are the same after we apply some permutation π of X to one of them: $P = \pi(Q)$. Clearly $\equiv$ is an equivalence relation on $\mathbb{P}(X)$. Let $[P]$ denote the equivalence class of P. See §I.A.4 (p. 45) for a review.

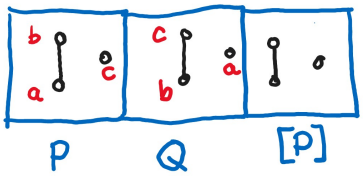


Figure 5: (a) $H(P)$; (b) $H(Q)$; (c) $[P]$ $(= [Q])$.
Here, $X = \{a, b, c\}$, $P = \{a < b\}$, $Q = \{b < c\}$.

Visual representations: the standard visualization of a partial order P is to draw its Hasse diagram (see Figure 5(a)). The Hasse diagram is a DAG whose nodes are labeled by the elements of X. Edges in the DAG are (implicitly) directed downwards, with the larger element above the smaller element. By the **Hasse shape** of P, we mean the Hasse diagram of P in which the labels of nodes are omitted. See Figure 5(c). Clearly all equivalent partial orders have the same Hasse shape.

(i) Draw all the 5 Hasse shapes of $\mathbb{P}(X)$ when $|X| = 3$.

(ii) Comparison tree program T can be quite large and cumbersome even for small values of n. We can compactly represent T by drawing its **Hasse tree program** $H(T)$. For example, if T is the tree program in §I.4, Figure 3 (page 9), then $H(T)$ is shown in Figure 6.

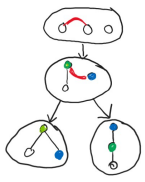


Figure 6: Hasse tree program for finding the maximum of 3 elements.

We can define $H(T)$ as follows:
(0) $H(T)$ is a rooted tree whose nodes have either one or two children.

(1) Each node u of $H(T)$ is represented by a Hasse shape denoted $H(u)$.
(2) At the root u_0, the Hasse shape $H(u_0)$ is that of the trivial partial order, $P = \varnothing$.
(3) If u is a non-leaf, we also indicate a single comparison: this is done by drawing a red (or dashed) edge between two nodes of $H(u)$.
(4) The non-leaf u has one or two children, depending on whether the two possible outcomes of the comparison at u produces one or two distinct Hasse shapes. Note that the root u_0 always has only one child.

Please design a Hasse tree program $H(T)$ to compute the second largest of $|X| = 4$ elements. It should have height 4.

(iii) The previous exercise to find the second largest of 4 elements can be generalized to finding the second largest of n elements:
STEP 1: find the largest with the help of a binary tree T_n with n leaves and height $\lceil \lg n \rceil$ as follows: the n input elements are placed at the leaves of T_n, and compare $x : y$ for each pair x, y of siblings. The larger of x, y is now placed at node of their parent. Continue this level-by-level comparison until the largest element is placed at the root of T_n. Note that $\lg = \log_2$ in this course.
STEP 2: find the largest among the contenders for the second place. Return this element.

Please answer 3 questions:
(a) Who are the contenders for the second place in STEP 2?
(b) What is the worst case number of comparisons in this algorithm?
(c) Illustrate the algorithm by drawing a Hasse tree program for the case $n = 8$.

♢

Exercise 4.4: Design tree programs for four elements a, b, c, d:
(a) To find the second largest element. The height of your tree should be 4 (the optimum).
(b) To sort the four elements. The height of your tree should be 5 (the optimum).

♢

Exercise 4.5: We are interested in comparator circuits $C_n(x_1, \ldots, x_n)$ to compute the maximum and minimum of $(x_1, \ldots, x_n)$. In other words, if output is (y_1, y_2) then $y_1 = \max\{x_1, \ldots, x_n\}$ and $y_2 = \min\{x_1, \ldots, x_n\}$. Draw the circuit for $n = 8$. Tell us how many comparators is needed in your C_8. ♢

Exercise 4.6: (a) Show that the median of 4 elements can be computed with 4 comparisons in the worst case. (b) Show that the median of 5 elements can be computed with 6 comparisons in the worst case. HINT: you could use your solution in part(a) for this part.
NOTE: the median of a set X of elements is the element of rank $\lceil |X|/2 \rceil$. An element has rank k if it is smaller than or equal to $k - 1$ other elements and larger than or equal to $|X| - k$ other elements. Thus rank 1 is the largest, and rank $|X|$ is the smallest element. In case the elements are non-distinct, an element could have several ranks. For instance, if all the elements are identical, then each element could have ranks 1 to $|S|$. ♢

Exercise 4.7: We want to determine the median of 5 elements using comparison trees. Figure 7 shows two Hasse diagrams on 5 elements. The partial order represented by diagram H_0 shows that c is the median.

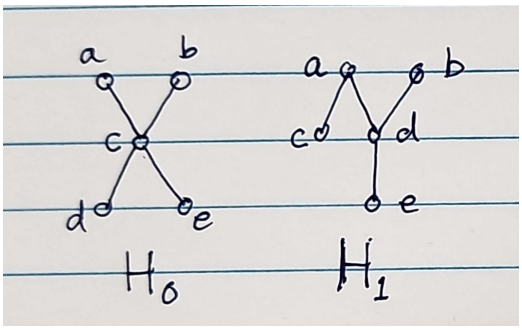
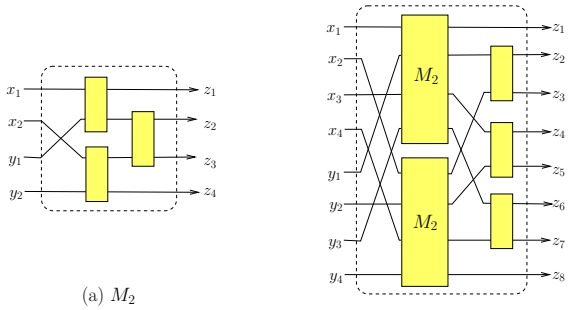


Figure 7: Hasse diagrams H_0 and H_1

Suppose that the Hasse diagram at some node in your comparison tree is H_1. In the worst case, how many more comparisons do you need to determine the median? Justify. ♢



(a) M_2 (b) M_4

Figure 8: Circuits M_2 and M_4

Exercise 4.8: Prove or disprove: the circuits M_2 and M_4 in Figure 8 correctly merges their inputs. ♢

Exercise 4.9: Assume n is a power of 2.
(a) Recursively describe a circuit S_n for sorting n numbers. You may use merging circuits M_k for any $k \leqslant n/2$ as subcircuits. Also, please draw the comparator circuit S_8. (b) Define the **delay** of a circuit to be the maximum number of comparators in a path from input node to output node. Suppose the maximum delay of the circuit M_n in part(a) is $\lg^2 n$. What is the delay of S_n in your construction (a)? ♢

Exercise 4.10: Prove the correctness of the circuits B_n in the text for merging two sorted lists of n numbers each. Use the "zero-one principle": *if a circuit correctly merges any* $0-1$ *inputs, then the circuit is correct.* ♢

Exercise 4.11: Design a tree program to merge two sorted lists (x, y, z) and (a, b, c, d). The height of your tree should be 6. Can you prove that 6 is optimum? ♢

Exercise 4.12: ("Unrolling" or "unwinding" a uniform algorithm into a comparison tree)
(a) Draw the comparison tree $T_{2,4}$ obtained by unrolling our Merge Algorithm on input (x_1, x_2) and (y_1, y_2, y_3, y_4).
(b) Draw the Hasse diagram associated with each node of the tree in Part (a). ♢

Exercise 4.13: Design a Tape Algorithm for sorting. Assume that the input tape is T_0 containing a sequence of n items. Finally, you must output the sorted items in tape T_0. Besides T_0, it is sufficient to have two other tapes T_1 and T_2, but you can have any number of additional tapes if you like. Our goal is to minimize the number of `RESET`'s ($O(\log n)$ suffices).

HINT: Use some form of Merge Sort. Use pseudo-code as we have done in the text to describe the Tape Algorithm for Merging. One key concept is the notion of a **run** which is any longest contiguous sequence of items in a tape that is increasing. E.g., $(1, 8, 2, 5, 9, 4)$ has three runs: $(1, 8), (2, 5, 9), (4)$. You want to merge the runs. But first you need to "distribute" the runs in the input tape into two other tapes, and then merge them: call this procedure $Distribute(T_0, T_1, T_2)$. ♢

Exercise 4.14: In the text, we described a procedure called $TapeMerge(T_1, T_2, T_0)$ to merge items in T_1 and in T_2 into T_0, *assuming* that the items in T_1, T_2 are already sorted. Let us explore what happens in case T_1 and T_2 have *many* runs. In the previous exercise, you also designed a procedure $Distribute(T_0, T_1, T_2)$ to distribute the runs in T_0 into T_1 and T_2 (alternately). For this question, assume that $Distribute(T_0, T_1, T_2)$ returns the number $k \geqslant 1$ of runs that was originally in T_0. Consider the following algorithm:

```
Input: Items in T_0. Initially T_1, T_2 empty.
  k ← Distribute(T_0, T_1, T_2).
  While (k > 1)
      TapeMerge(T_1, T_2, T_0)
      k ← Distribute(T_0, T_1, T_2)
```

Clearly, using $TapeMerge$ in this way is an abuse of our original intention. *But if this algorithm halts*, it would have successfully sorted the items in the original tape T_0. Our goal is to prove that it will, in fact, halt. REMARK: this algorithm was accidentally discovered by students when I asked them to design a tape sorting algorithm from scratch.

HINT: Consider the runs that appear in T_0 after $TapeMerge$. When does a new run appear in T_0? Can you bound the number of such events? ♢

Exercise 4.15: How do you speed up your algorithm in the previous exercise if you have 4 tapes? Note that "reducing speed" here means using fewer `RESET`'s in your algorithm. What if you have $k > 4$ tapes? ♢

Exercise 4.16: In the tape model, it is non-trivial to reverse the contents of a tape. For instance, if the input tape contains (a, b, c, d), we want the output tape to contain (d, c, b, a). Give an $O(\log n)$ pass algorithm to reverse a list of n items in a tape. HINT: the method is very similar to the tape merge or tape sort algorithm. ♢

End Exercises

§5. Complexity Model: How to assess algorithms?

We now have a suitable computational model for solving our problem. What is the criteria to choose among different algorithms within a model? For this, we need to introduce a **complexity model**.

In most computational models, there are usually natural notions of **time** and **space**. These are two examples of **computational resources**. We view resources as a scarcity, and algorithms consume resources when they run. We want to design algorithms that minimize the use of resources. For this purpose, we shall focus on an algorithm's usage of only one resource, ignoring its behavior on the other resources. This resource is usually the time (occasionally space) resource. Thus we avoid studying the simultaneous usage of two or more resources, as this involves the more delicate issue of trade-offs between resources.

Next, for each primitive operation executing on a particular data, we need to know how much of the time resource is consumed. For instance, in `Java`, we could define each execution of the addition operation on two numbers a, b to use time $\log(|a| + |b|)$. Or again, the comparison $a : b$ of two integers in the comparison tree model may be charged $\log(|a| + |b|)$. But it would be simpler to say that these operation takes unit time, independent of a, b. This simpler version is our choice throughout these lectures: *each primitive operation takes unit time, independent of the actual arguments to the operation.*

Let us return to our 3 models for sorting-like problems: how do we assess algorithms in each model?

- In the Comparison Tree Model, counting each comparison as taking unit time, we conclude that the worst case time is the *height* of the comparison tree. Thus, we assess such programs in terms of their height. This complexity measure seems quite natural.
- In the Comparator Circuit Model, we could count the number of comparators in the circuit (this is called the **circuit size**). Evidently, circuit size is not a measure of "time" – so where is time complexity here? See ¶22 below. In general, the circuit model is a form of parallel computation model. Parallelism is an important topic that is outside the scope of this book.
- The Tape Model will turn out to be even less familiar. The key is to realize that among the 6 primitives, the `RESET` primitive is by far the most expensive. Physically, it means we must unwind the entire tape onto another (initially empty) spool. Therefore we seek tap algorithms that minimizes the number of resets. For instance, our Tape Merge Algorithm uses 3 `RESET`'s. If we assume all tape heads are initially at position 0, then the algorithm uses no `RESET`'s.

After assigning a (time) cost to each primitive operation, for each algorithm A, and for each input instance x, we could now assign a number $T_A(x)$ which is the (time) complexity of algorithm A on input x. If X is the input domain for A, this defines the corresponding complexity function

$$T_A : X \to \mathbb{R}. \tag{8}$$

If B is another algorithm on domain X, we also have $T_B : X \to \mathbb{R}$. Thus allows us to compare A and B by comparing T_A and T_B. E.g., A is at least as good as B if for all $x \in X$, $T_A(x) \leqslant T_B(x)$. But to develop a complexity theory, we want a way to discuss the complexity of algorithm A without reference to other algorithms.

Usually, the input domain X has infinitely many elements, and there is a notion of "input size", as given by the function (5). Now we can measure resource usage as a function of input size, using the following procedure. We define the **worst case complexity** of A to be the function $T_A : \mathbb{N} \to \mathbb{R}$ where

$$T_A(n) := \sup\{T_A(x) : x \in X, size(x) = n\} \tag{9}$$

Using sup illustrates one way to "aggregate" the set $\{T_A(x) : x \in X, size(x) = n\}$ of numbers. Instead of sup, we can also take the average to get **average complexity**. In general, we may apply any aggregating function G and define

$$T_A(n) = G(\{T_A(x) : x \in X, size(x) = n\}). \tag{10}$$

In non-uniform models such as tree programs or comparator circuits, we may take A to a non-uniform program as in (6).

In summary, a **complexity model** is a specification of
(a) The computational resource (e.g., time, space),
(b) The cost (in terms of the computational resource) of primitive operations (e.g., unit cost, logarithmic cost),
(c) The input size function, $size : X \to \mathbb{N}$, and
(d) The method G of aggregating (e.g., worst case, average).
Once the complexity model is fixed, we can associate to each algorithm A a **complexity function**

$$T_A : \mathbb{N} \to \mathbb{R}. \tag{11}$$

We cannot overstate the theoretical advantage of the function (11) over (8). (achieved with the help of (9)). We can now compare the complexity of two different problems even when their input domains are different. Complexity theory is founded[7] on functions such as (11). Note that (11) is possible thanks to the size function (5) and the aggregation method (10).

¶19. From Complexity of Algorithms to Complexity of Problems. The complexity function T_A concerns a single algorithm A. But properly speaking, *Complexity Theory is the study of the complexity of problems, not of algorithms per se.* If P is a problem, we need to consider the set of *all* T_A where A ranges over all algorithms A for P. Naturally, the A's must be programs in a fixed computational model. Among all these algorithms, it would be nice if there exists an algorithm A^* for P whose complexity T_{A^*} is "optimal". In general, optimal algorithms may not exist. But for non-uniform algorithms, we can always define the optimal algorithm $A^* = \{A^*_n : n \in \mathbb{N}\}$ since each A^*_n can be chosen to have minimal height. The complexity of such an optimal algorithm may be called the **inherent complexity** of the problem P (relative to the computational model). We next introduce two such examples.

¶20. Example (T1). Inherent Complexity of Sorting. Consider the comparison tree model for sorting. For each $n \in \mathbb{N}$, let $S(n) := \min_A T_A$ where A ranges over all comparison trees that sort n elements (recall that T_A is the height of the tree A). The function $S : \mathbb{N} \to \mathbb{N}$ so defined is called the **inherent complexity of sorting**. For instance, it is easy to see that $S(1) = 0$ and $S(2) = 1$. It is "inherent" because it is not a function of a single algorithm, but speaks to all possible algorithms for sorting. This is because even uniform algorithms can be unrolled into a non-uniform comparison tree algorithm.

[7] In situations where there is no suitable size functions, e.g., for $X = \mathbb{R}$, only an impoverished complexity theory can be developed.

We now prove our first non-trivial result in this chapter. Start with the simple observation: *any tree program A to sort n elements must have at least $n!$ leaves.* This is because A must have at least one leaf for each possible sorting outcome, and there are $n!$ outcomes when the input elements are all distinct. But a binary tree A of height h has at most 2^h leaves. Hence $2^h \geqslant n!$ or $h \geqslant \lg(n!)$. This proves:

Lemma 1 (Information-Theoretic Bound) *Every tree program for sorting n elements has height at least $\lg(n!)$, i.e.,*

$$S(n) \geqslant \lg(n!). \tag{12}$$

This result is called the **Information Theoretic Bound** (ITB) for sorting. To use (12) effectively, it helpful to know $\lg(n!)$ is roughly $n \lg n$ (see Exercises for more precise formulations). Thus, it tells use that you need at least "$n \lg n$" comparisons to sort n elements.

Since $S(n)$ is an integer, the ITB bound (12) is equivalently to $S(n) \geqslant \lceil \lg(n!) \rceil$. Defining $\texttt{ITB}(n) := \lceil \lg(n!) \rceil$, here are its first 10 values:

n:	1	2	3	4	5	6	7	8	9	10
$\texttt{ITB}(n)$:	0	1	3	5	7	10	13	16	19	22

E.g., you need at least 22 comparisons to sort 10 elements. This deceptively simple result is really quite deep: to appreciate this, try proving by direct arguments that, in the worst case, you need more than four comparisons to sort four elements. The wise words of Fraleigh[8] (see margin) extend to inequalities like the ITB. How good is the ITB lower bound on $S(n)$? Trivially, we can see that $S(1) = 0$ and $S(2) = 1$. What about $n = 3$? By ITB, $S(3) \geqslant \lg 3!$, so $S(3) \geqslant \lceil \lg 6 \rceil = 3$. Let us check the next simplest case, $S(3)$. It is easy to see that you can sort three elements in at most 3 comparisons: if you are given distinct x, y, z then you can begin by comparing $x : y$ and $x : z$. If you are lucky, this might end up sorting the elements (either $y > x > z$ or $z > x > y$). Otherwise one more comparison $y : z$ will sort the input. *This is obvious if you use Hasse diagrams!* This proves $S(3) \leqslant 3$. Combined with the ITB, we conclude that $S(3) = 3$, and so the ITB bound is tight.

"Never underestimate a theorem that counts something"
– J.B. Fraleigh

Note how the proof of $S(3) = 3$ requires two distinct arguments: an upper bound argument $S(3) \leqslant 3$ amounts to providing an algorithm. The lower bound argument $S(3) \geqslant 3$ comes from ITB. In one sense, this simple problem illustrates the core of complexity theory: getting good upper (by devising algorithms) and lower bounds (by devising impossibility arguments) on computational problems. It is known that the ITB bound is optimal for $n \leqslant 31$.

Complexity Theory in a nutshell!

Open problem: determine $S(32)$

¶21. Example (T2). Inherent Complexity of Merging. We similarly define the **inherent complexity of merging** to be the function $M : \mathbb{N}^2 \to \mathbb{N}$ where $M(m, n)$ is the minimum height of any comparison tree for merging two sorted lists of sizes m and n, respectively. Let us prove the following upper and lower bounds:

[8] Fraleigh was referring to Lagrange's theorem on finite groups in **A First Course in Abstract Algebra**, Addison-Wesley 1969, p. 93, and other similar counting theorems in algebra. Since learning these words as an undergraduate, its wisdom has only grown on me over time.

Lemma 2 (Inherent Bounds on Merging)

$$(a) \quad M(m,n) \leqslant m+n-1, \tag{13}$$
$$(b) \quad M(m,n) \geqslant 2\min\{m,n\} - \delta[m=n], \tag{14}$$
$$(c) \quad M(m,n) \geqslant \lg \binom{m+n}{m}. \tag{15}$$

where $\delta[P] = 1$ if the predicate P is true, and $\delta[P] = 0$ otherwise.

Proof. The upper bound (13) comes from the Merge Algorithm in ¶14. The idea of the proof is to associate one comparison for each one output in the main loop. More formally, we devise a simple **charging scheme** whereby each comparison that the algorithm makes is "charged" to the element that is output as a result of the comparison. But you cannot charge more than the number of output elements. This gives an upper bound of $\leqslant m+n$ comparisons. We improve this bound by observing that the last element can be output without any comparison. Hence we obtain the sharper upper bound of $m+n-1$. This charging argument is a very elementary example of what we call an **amortized analysis** in Chapter 6.

The lower bound (14) comes from the following input instance: assume the input is $x_1 < x_2 < \cdots < x_m$ and $y_1 < \cdots < y_n$ where $m \geqslant n$ and

$$x_1 < y_1 < x_2 < y_2 < x_3 < \cdots < x_n < y_n (< x_{n+1} < \cdots < x_m).$$

Let us rename the first $2n$ elements as

$$z_1 < z_2 < z_3 < z_4 < z_5 < \cdots < z_{2n-1} < z_{2n}$$

where $z_{2i-1} = x_i$ and $z_{2i} = y_i$ $(i = 1, \ldots, n)$. We claim that the comparison $z_i : z_{i+1}$ must be made for each $i = 1, \ldots, 2n-1$.

> The claim follows from a simple fact about partial orders (see Appendix for definition). A relationship $x < y$ in a partial order P is **essential** if it cannot be deduced from other relationships in P. In the comparison model, every essential relationship must be determined by a comparison. In particular, the relationships $z_i < z_{i+1}$ are essential. These primitive relationships constitute the edges of a directed graph $H(P)$ called the **Hasse diagram** of P. Hasse diagrams are compact representations of partial orders since $H(P)$ has at most $n-1$ edges if P has n elements. E.g., is P is a linear order on n elements, then $H(P)$ is just a linear graph.

This yields a lower bound of $2n-1$ comparisons. In case $m > n$, there is at least one more comparison to be made, between y_n and x_{n+1}. So if $m > n$, we need at least $2n$ comparisons. This proves $M(m,n) \geqslant 2n - \delta(m,n)$, where $n = \min\{m,n\}$. This method of proving lower bounds is simple form of what are called **adversary arguments** in Chapter XII, where you imagine a 2-player game between the algorithm and an adversary.

Finally, the lower bound (15) is the **information-theoretic bound** (ITB) for merging (analogous to (12) for sorting). In proof, there are $\binom{m+n}{n}$ ways of merging the two sorted lists. To see this, note that each sorted list of $m+n$ elements is uniquely determined once we know the m positions that are filled by elements that come from the list of size m. There are $\binom{m+n}{m}$

ways of choosing these positions. This concludes our proof of Lemma 2. **Q.E.D.**

As corollary of the upper and lower bounds, we obtain some exact bounds for the complexity of merging:

$$M(m,m) = 2m-1$$

and

$$M(m,m+1) = 2m.$$

Thus the uniform algorithm is optimal in these cases. More generally, $M(m,m+k) = 2m+k-1$ for $k = 0,\ldots,4$ and $m \geqslant 6$ (see [5] and Exercise). These bounds are for inputs where $|m-n|$ is a small constant. Now consider the other extreme situation where $|m-n|$ is as large as possible: $M(1,n)$. In this case, the information theoretic bound says that $M(1,n) \geqslant \lceil \lg(n+1) \rceil$ (why?). Also, by binary search, this lower bound is tight (Exercise). Hence we now know another exact value:

$$M(1,n) = \lceil \lg(n+1) \rceil . \tag{16}$$

A non-trivial result from Hwang and Lin says

$$M(2,n) = \lceil \lg 7(n+1)/12 \rceil + \lceil \lg 14(n+1)/17 \rceil .$$

Thus we have two distinct methods for proving lower bounds on $M(m,n)$: the adversary method is better when $|m-n|$ is small, and the information theoretic bound is better when this gap is large. The exact value of $M(m,n)$ is known for several other cases, but a complete description of this complexity function remains an open problem.

Student: *I thought lower bounds on $S(n)$ is about the "best case" complexity of sorting*

¶* 22. Best Case Complexity. Although our main interest is in worst-case complexity, it is useful to briefly consider the notion of "best case complexity". Note that contrary to what some students think, lower bounds on $S(n)$ is *still* about worst case complexity, not about best case complexity.

Again, if A_n is a tree program that accepts inputs of size n, we define the **best case complexity** of A_n to be the length of the *shortest path* from the root of A_n to a leaf. We can apply this concept to sorting and to merging. Define $S'(n)$ to be the best case complexity for sorting n elements: $S'(n) := \min\{T'_{A_n}\}$ where A_n range over all tree programs that sort n elements. Similarly, define $M'(m,n)$ to be the best case complexity for merging m elements with n elements. We claim:

Student: *I thought $M'(m,n) = \min\{m,n\} - \delta(m,n)$*

$$S'(n) = n-1, \qquad M'(m,n) = 1. \tag{17}$$

To see $S'(n) = n-1$, we see that if the output of sorting is $(z_1,\ldots,z_n)$, then we must have made the comparisons $z_i : z_{i+1}$ for $i = 1,\ldots,n-1$. To see that $M'(m,n) = 1$, note that on input $(x_1,\ldots,x_m)$ and $(y_1,\ldots,y_n)$, we may be able to get away with a single comparison $x_m : y_1$.

¶* 23. Complexity of Comparator Circuits. Let C be a comparator circuit. There are two complexity measures of interest, **delay** $D(C)$ and **size** $S(C)$: $D(C)$ is the length of the longest path from an input node to an output node, and $S(C)$ is the number of comparators in C. Since values move in parallel, $D(C)$ can be thought off as "parallel time". In 1968, Batcher introduced circuits (see Exercise) for sorting n numbers with delay $\mathcal{O}(\log^2 n)$ and size $\mathcal{O}(n \log^2 n)$. In 1983, Ajtai, Komlós and Szemerédi made a major breakthrough by proving

the existence of sorting circuits with delay $\mathcal{O}(\log n)$ and size $\mathcal{O}(n \log n)$. This is known as the **AKS sorting circuit**. It is an understatement to say that the AKS circuit is impractical.[9] In 2004, Goodrich made another breakthrough by providing a sorting circuit of size $O(n \log n)$ that seems practical. We return to this topic in Chapter 12.

¶* 24. Other Complexity Measures. We briefly look at some other kinds of complexity measures.

- In computational geometry, it is often useful to take the output size into account. The complexity function would now take at least two arguments, $T(n, k)$ where n is the input size, but k is the output size. This is the **output-sensitive complexity model**.
- Another kind of complexity measure is the **size** of a program. In the RAM model, this can be the number of primitive instructions. We can measure the complexity of a problem P in terms of the size $s(P)$ of the smallest program that solves P. This complexity measure assigns a single number $s(P)$, not a complexity function, to P. This **program size measure** is an instance of **static complexity measure**; in contrast, time and space are examples of **dynamic complexity measures**. Here "dynamic" ("static") refers to fact that the measure depends (does not depend) on the running of a program. Complexity theory is mostly developed for dynamic complexity measures.
- The comparison tree complexity model ignores all the other computational costs except comparisons. In most situations this is well-justified. But it is possible[10] to conjure up ridiculous algorithms which minimize the comparison cost, at an exorbitant cost in other operations.
- The size measure is relative to representation. Perhaps the key property of size measures is that *there are only finitely many objects up to any given size*. Without this, we cannot develop any complexity theory. If the input set are real numbers, $\mathbb{R}$, then it is very hard to give a suitable size function with this property. This is the puzzle of real computation.

Exercises

Exercise 5.1: How many comparisons are required in the worst case to sort 10 elements in the comparison tree model? In other words, give a lower bound on $S(10)$. HINT: to do this computation by hand, it is handy to know that $10! = 3,628,800$ and $2^{20} = 1,048,576$. ♢

Exercise 5.2: Show that $\lg(n!) \geqslant (n/2)\lg(n/2)$. HINT: use the fact that $\lg(ab) = \lg(a) + \lg(b)$ to write $\lg(n!)$ as a summation. Then discard the smallest $n/2$ elements. Since $(n/2)$ might not be an integer, use the relation $n = \lfloor n/2 \rfloor + \lceil n/2 \rceil$. ♢

[9] Knuth noted that unless n exceeds the number of atoms in the universe, the size of Batcher's circuit is smaller than AKS circuit. In the blog of Lipton, such algorithms are called " galactic algorithms".

[10] My late colleague, Professor Robert Dewar gave the following example: given n numbers to be sorted, first search for all potential comparison trees for sorting n elements. To make this search finite, only evaluate comparison trees of height at most $n\lceil \lg n \rceil$. Among those trees that we have determined to be able to sort, pick one of minimum height. Now run this comparison tree on the given input.

Exercise 5.3: How good is the information theoretic lower bound? In other words, can we find upper bounds that match information-theoretic lower bound? We know it is tight for $S(3)$. What about $S(4)$? ♢

Exercise 5.4: (a) Give a lower bound for $S(5)$.
(b) Show that if the first two comparisons should involve only three input elements, then you need a total of at least 8 comparisons in the worst case.
(c) Using the insight from (b), give an optimal algorithm to sort 5 elements. ♢

Exercise 5.5: The exact value of $S(1000)$ is unknown, so we seek to derive upper and lower bounds on $S(1000)$. For upper bound, please bound the complexity of any $O(n \log n)$ algorithm, say, MergeSort.

NOTE: we are asking for two numbers. State the two values and tell us how you obtain them. Your numbers must be *explicit* (in decimal notation like 1234), not an expression like $1000^2\,\lceil \lg 1000 \rceil$. You may use computer programs or calculators, etc, but tell us how you do it. If you use computers, discuss how you can be confident despite rounding errors in the computation. Stirling's formula might be useful to know.

♢

Exercise 5.6: Give upper and lower bounds for $S(100)$.
HINT: Unlike the previous question for $S(1000)$, we suggest better upper bounds for $S(100)$ by exploiting the fact that $S(n)$ is known exactly for all $n \leqslant 31$, and it is achieved by the ITB bound. In particular, use the formula:

$$S(100) \leqslant \big(S(30) + S(30) + M(30;30)\big) + \big(S(30) + S(10) + M(30;10)\big) + M(60;40).$$

♢

Exercise 5.7: The following is a variant of the previous exercise. Is it always possible to sort n elements using a comparison tree with $n!$ leaves?
(a) Let $L(n)$ be the minimal number of leaves in a comparison tree that sorts n elements. Clearly, $L(n) \geqslant n!$. Show that $L(n) = n!$.
(b) Let us make the problem a bit harder: let $L^*(n)$ be the minimal number of leaves in a comparison tree that sorts n elements in (optimal) height $S(n)$. Clearly $L^*(n) \geqslant L(n) = n!$. Give the best upper bound you can for $L^*(n)$ when $n = 3, 4, 5$. ♢

Exercise 5.8: Suppose we can sort 10 elements optimally. Here is a proposed algorithm to sort 20 elements: divide the 20 elements into two equal size groups, and sort each group optimally; return the merge of the two sorted lists.

State a condition (which can be verified with a numerical computation which you need not do) that is equivalent to this algorithm being optimal.
HINT: the information theoretic bound (ITB) is optimal for $n \leqslant 31$.

♢

Exercise 5.9: (a) Consider a variant of the unit time complexity model for the integer RAM model, called the **logarithmic time complexity model**. Each operand takes time that is logarithmic in the address of the register and logarithmic in the size of its operands. What is the relation between the logarithmic time and the unit time models?
(b) Is this model realistic in the presence of the arithmetic operators (ADD, SUB, MUL, DIV). Discuss. ♢

Exercise 5.10: Describe suitable complexity models for the "space" resource in integer RAM models. Give two versions, analogous to the unit time and logarithmic time versions. What about real RAM models? ♢

Exercise 5.11: Prove the claim (16) that $M(1,n) \leqslant \lceil \lg(n+1) \rceil$. ♢

Exercise 5.12: Give your best upper and lower bounds for $M(2,10)$. For upper bound, please give an explicit method. ♢

Exercise 5.13: Prove that $M(m, m+i) = 2m + i - 1$ for $i = 2, 3, 4$ for $m \geqslant 6$. ♢

Exercise 5.14: Prove that $M(k,m) \geqslant k \lg_2(m/k)$ for $k \leqslant m$. HINT: split the list of length m into three sublists of roughly equal sizes. ♢

Exercise 5.15: Open problem: determine $M(m,3)$ and $M(m, m+5)$ for all m. ♢

Exercise 5.16: Let C be a comparator circuit for a problem P. If C has h comparators, show that there exists is a comparison tree A with height h that solves problem P. You must argue in what sense P is solved. ♢

Exercise 5.17: Describe time and space complexity models for the comparator circuit model in ¶9. Then define $T(n)$ and $S(n)$ as the inherent time and inherent space to sort n numbers in this model.
HINT: "Time" is the maximum number of comparisons along any path, and "space" is the number of comparators. Derive bounds on $T(n)$ and $S(n)$. ♢

Exercise 5.18: Suppose $X_1, \ldots, X_n$ are n sorted lists, each with k elements. Show that the complexity of sorting the set $X = \bigcup_{i=1}^{n} X_i$ is $\Theta(nk \log n)$. ♢

End Exercises

§6. Algorithmic Techniques: How to design efficient algorithms

Now that we have some criteria to judge algorithms, we begin to design algorithms that are "efficient" according to such criteria. There emerges some general paradigms of algorithms design:
(i) Divide-and-conquer (e.g., merge sort)
(ii) Greedy method (e.g., Kruskal's algorithm for minimum spanning tree)
(iii) Dynamic programming (e.g., multiplying a sequence of matrices)
(iv) Incremental method (e.g., insertion sort)

Let us briefly outline the merge sort algorithm to illustrate the divide-and-conquer paradigm: Suppose you want to sort an array A of n elements. Here is the **Merge Sort** (or Mergesort) algorithm on input A:

Merge Sort Algorithm
Input: An array A with $n \geqslant 1$ numbers.
Output: The sorted array A containing these numbers, but in increasing order.
0. (Basis) If $n = 1$, return the array A.
1. (Divide) Divide the elements of A into two subarrays B and C of sizes $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$ each.
2. (Recurse) Recursively, call the Merge Sort algorithm on B. Do the same for C.
3. (Conquer) Merge the sorted arrays B and C into the array A

It is important to note that this is a recursive algorithm: the algorithm calls itself on smaller size inputs. E.g., to Merge Sort (a, b, c, d), you have to recursively Merge Sort (a, b) and (c, d). Besides recursion, there is only one non-trivial step in this algorithm, the Conquer Step which merges two sorted arrays. The subalgorithm for merging was already present in ¶14.

There are many variations or refinements of these paradigms. E.g., Kirkpatrick and Seidel [4] introduced a form of divide-and-conquer (called "marriage-before-dividing") that leads to an output-sensitive convex hull algorithm. There may be domain specific versions of these methods. E.g., plane sweep is an incremental method suitable for problems on points in Euclidean space.

Closely allied with the choice of algorithmic technique is the choice of *data structures*. A data structure is a representation of a complex mathematical structure (such as sets, graphs or matrices), together with algorithms to support certain querying or updating operations. For instance, to implement recursive algorithms such as Merge Sort above, we will need the use of a "stack" to organize the recursive calls. A stack is an example of a basic data structure. The following are a list of such basic data structures.

(a) Linked lists: each list stores a sequence of objects together with operations for (i) accessing the first object, (ii) accessing the next object, (iii) inserting a new object after a given object, and (iv) deleting any object.

(b) LIFO, FIFO queues: each queue stores a set of objects under operations for insertion and deletion of objects. The queue discipline specifies which object is to be deleted. There are two[11] basic disciplines: last-in first-out (LIFO) or first-in first-out (FIFO). Note that recursion is intimately related to LIFO.

(c) Binary search trees: each tree stores a set of elements from a linear ordering together with the operations to determine the smallest element in the set larger than a given element. A dynamic binary search tree supports, in addition, the insertion and deletion of elements.

(d) Dictionaries: each dictionary stores a set of elements and supports the operations of (i) inserting a new element into the set, (ii) deleting an element, and (iii) testing if a given element is a member of the set.

[11] A discipline of a different sort is called GIGO, or, garbage-in garbage-out. This is really a law of nature.

(e) Priority queues: each queue stores a set of elements from a linear ordering together with the operations to (i) insert a new element, (ii) delete the minimum element, and (iii) return the minimum element (without removing it from the set).

Exercises

Exercise 6.1: (a) Design an incremental sorting algorithm based on the following principle: assuming that the first m elements have been sorted, try to add ("insert") the $m+1$st element into the first m elements to extend the inductive hypothesis. Moreover, assume that you do all these operations using only the space in the original input array.

(b) If the number n of elements to be sorted is small (say $n < C$), this approach can lead to a sorting algorithm that is faster than Merge Sort. Intuitively it is because Merge Sort uses recursion that has non-trivial overhead cost. So a practical implementation of Merge Sort might switch an incremental sorting method as in part(a) when $n < C$. Design such a hybrid algorithm that combines the Merge Sort algorithm with your solution in (a).

(c) Implement the Merge Sort Algorithm, your incremental sorting algorithm of part(a), and the hybrid algorithm in part(b). Try to see if you can experimentally verify our remarks in (b), and determine the constant C. ♢

End Exercises

§7. Analysis: How to estimate complexity

We have now a measure T_A of the complexity of our algorithm A, relative to some complexity model. Unfortunately, the function T_A is generally too complex to admit a simple description, or to be expressed in terms of familiar mathematical functions. Instead, we aim to give upper and lower bounds on T_A. This constitutes the subject of **algorithmic analysis** which is a major part of this book. The tools for this analysis depend to a large extent on the algorithmic paradigm or data structure used by A. We give two examples.

¶25. Example (D1) Divide-and-Conquer. If we use divide-and-conquer then it is likely we need to solve some recurrence equations. In our Merge Sort algorithm, assuming n is a power of 2, we obtain the following recurrence:

$$T(n) = 2T(n/2) + Cn$$

for $n \geqslant 2$ and $T(1) = 1$, and $C \geqslant 1$ is some constant determined by the complexity of merging. Here $T(n) = T_A(n)$ is the (worst case) number of comparisons needed by our algorithm A to sort n elements. The solution is $T(n) = \Theta(n \log n)$. In the next chapter, we study techniques to obtain such solutions.

¶26. Example (D2) Amortization. If we employ certain data-structures that might be described as "lazy" then amortization analysis might be needed. Let us illustrate this with the problem of maintaining a binary search tree under repeated insertion and deletion of elements.

Ideally, we want the binary tree to have height $\mathcal{O}(\log n)$ if there are n elements in the tree. There are a number of known solutions for this problem (see Chapter 3). Such a solution achieves the optimal logarithmic complexity for *each* insertion/deletion operation. But it may be advantageous to be lazy about maintaining this logarithmic depth property: such laziness may be rewarded by a simpler coding or programming effort. The price for laziness is that our complexity may be linear for individual operations, but we may still hope to achieve logarithmic cost in an "amortized" sense (thought of as a kind of averaging). To illustrate this idea, suppose we allow the tree to grow to non-logarithmic depth as long as it does not cost us anything (*i.e.*, there are no queries on a leaf with big depth). But when we have to answer a query on a "deep leaf", we take this opportunity to restructure the tree so that the depth of this leaf is now reduced (say halved). Thus repeated queries to this leaf will make it shallow. The cost of a single query could be linear time, but we hope that over a long sequence of such queries, the cost is amortized (averaged) to something small (say logarithmic). This technique prevents an adversary from repeated querying of a "deep leaf". But how do we account for the first few queries into some "deep leaves" which have linear costs? To anticipate such expenses, the idea is to "pre-charge" those initial insertions that lead to this inordinate depth. Using a financial paradigm, we put the pre-paid charges into some bank account. Then the "deep queries" can be paid off by withdrawing from this account. Amortization is both an algorithmic paradigm as well as an analysis technique. This will be treated in Chapter 6.

§8. Asymptotics: How robust is the model?

This section contains important definitions for the rest of the book.

Take note!

Let us review what we have done so far: we started with a problem P (§3), selected an appropriate computational model (§4) and an associated complexity model (§5), and designed an algorithm A for P in our model (§6). As we next embark on the problem of analyzing the complexity function T_A of our algorithm (§7) in order to understand how good or efficient is our algorithm. We quickly realize that T_A will generally be too complex to determine exactly. But looking back at this process, we are bound to find many arbitrary choices that affects T_A. For instance, would a simple change in the set of primitive operations drastically change the complexity of your solution? Or what if we charge two units of time for some of the operations? Of course, there is no end to such revisionist afterthoughts. What we are really seeking is a certain robustness or invariance in our results. This section addresses this important concern.

¶27. Partial and total functions. Let $f : D \to R$ be a function, where D is called the **domain** and R the **range**. In ordinary discourse, this is understood to mean that for every $x \in D$, the function f returns a value $f(x) \in R$. We are now going to consider a slightly more general concept. We call $f : D \to R$ a **partial function** if for all $x \in D$, either $f(x)$ is either **defined**, in which case $f(x)$ represents an element of R, or else $f(x)$ is **undefined**, and does not represent anything. We shall write write $f(x) = \uparrow$ if $f(x)$ is undefined, and write $f(x) = \downarrow$ if it is defined. The partial function f is said to be a **total function** if for all $x \in D$, $f(x)$ is defined (and hence $f(x) \in R$). In other words, total functions are the kind of functions we ordinarily assume. But in the presence[12] of partial functions, we need to give it a name.

$\uparrow$ may be seen as a special value, but $\downarrow$ is only a surrogate for all other values!

[12] We remark that the literature sometimes use the notation $f : D \dashrightarrow R$ to indicate that f is a partial function. However, we shall not use this "$\dashrightarrow$" notation.

¶28. What is a complexity function? In this book, we call a partial function of the form

$$f : \mathbb{R}^n \to \mathbb{R}$$

a **complexity function**. Usually, we have $n = 1$. We use complexity functions to quantify the complexity of algorithms. Why do we consider *partial* functions for complexity functions? For one thing, many functions of interest are only defined on positive integers. For example, the running time $T_A(n)$ of an algorithm A that takes discrete inputs is a partial real function (normally defined only when n is a natural number). Of course, if the domain of T_A is taken to be $\mathbb{N}$, then $T_A(n)$ might perhaps be total. Still, we prefer to use $\mathbb{R}$ as the domain of $T_A(n)$. This is because we often use functions such $f(n) = n/2$ or $f(n) = \sqrt{n}$, to bound our complexity functions, and these are naturally defined on the real domain; all the tools of analysis and calculus become available to analyze such functions. Many common real functions such as $f(n) = 1/n$ or $f(n) = \log n$ are partial functions because $1/n$ is undefined at $n = 0$ and $\log n$ is undefined for $n \leqslant 0$.

We have to be careful about operations on partial functions, and when they are used to define predicates. We have a general rule for composition of two partial functions $f, g : \mathbb{R} \to \mathbb{R}$:

$$g(x) = \uparrow \text{ implies } f(g(x)) = \uparrow . \tag{18}$$

More simply: $f(\uparrow) = \uparrow$ is always true. In general, if any argument of a function is undefined, then the value of the function is undefined.

¶29. Partial Predicates. For any set D, a partial function $P : D \to \{0, 1\}$ is called a **partial predicate** over D. We say the predicate P **holds at** $x \in D$ if $P(x) = 1$. So 1 is the "true" value and 0 is the "false" value. The partial predicate P is **valid** if for all $x \in D$, either $P(x) = \uparrow$ or $P(x) = 1$. If $P(x) = \uparrow$ for all $x \in D$, then we say P is **vacuously** valid. Partial predicates arise naturally from relations among partial functions. If f, g are complexity functions, then the relation "$f \leqslant g$" represents the partial predicate $P : \mathbb{R} \to \{0, 1\}$ where $P(x) = \uparrow$ if $f(x) = \uparrow$ or $g(x) = \uparrow$; otherwise, $P(x) = \downarrow$. Naturally, when $P(x) = \downarrow$, we have $P(x) = 1$ iff $f(x) \leqslant g(x)$.

If $P_i : D \to \{0, 1\}$ are partial predicates ($i = 0, 1$), then so are $\neg P_i$, $P_0 \vee P_1$ and $P_0 \wedge P_1$ (recall our general rule (18) about composition of partial functions).

How to quantify over partial predicates

Quantification over partial predicates is defined as follows: The sentence "$(\forall x)P(x)$" holds iff for all $x \in D$, either $P(x) = \uparrow$ or $P(x) = 1$. Similarly "$(\exists x)P(x)$" holds iff there is some $x \in D$ such that $P(x) = \downarrow$ and $P(x) = 1$. E.g., if P is the always undefined predicate, then $(\exists x)P(x)$ is false. De Morgan's law for quantifiers say that

$$\left.\begin{array}{rcl} \neg(\forall x)P(x) & \equiv & (\exists x)\neg P(x). \\ \neg(\exists x)P(x) & \equiv & (\forall x)\neg P(x). \end{array}\right\} \tag{19}$$

It is not hard to see that (19) holds even when P is a partial predicate.

¶30. Designated variable and Anonymous functions. In general, we will write "n^2" and "$\log x$" to refer to the functions $f(n) = n^2$ or $g(x) = \log x$, respectively. Thus, the functions denoted n^2 or $\log x$ are **anonymous** (or, perhaps more accurately, "self-naming"). This convention is very convenient, but it relies on an understanding that "n" in n^2 or "x" in $\log x$ is the **designated variable** in the expression. For instance, the anonymous complexity function $2^x n$ is a linear function if n is the designated variable, but an exponential function if x

is the designated variable. *The designated variable in complexity functions, by definition, range over real numbers.* This may be a bit confusing when the designated variable is "n" since in mathematical literature, n is usually a natural number.

n might be a real variable!

¶31. Robustness or Invariance issue. Let us return to the robustness issue which motivated this section. The motivation was to state complexity results that have general validity, or independent of many apparently arbitrary choices in the process of deriving our results. There are many ways to achieve this: for instance, we can specify complexity functions up to "polynomial smearing". More precisely, two real functions f, g are said to be **polynomially equivalent** if for some $c > 0$, $f(n) \leqslant cg(n)^c$ and $g(n) \leqslant cf(n)^c$ for all n large enough. Thus, $\sqrt{n}$ and n^3 are polynomially equivalent according to this definition. This is *extremely* robust but alas, too coarse for most purposes. The most widely accepted procedure is to take two smaller steps:

two steps towards invariance

- Step 1: We are interested in the eventual behavior of functions. E.g., if $T(n) = 2^n$ for $n \leqslant 1000$ and $T(n) = n$ for $n > 1000$, then we want to regard $T(n)$ as a linear function.
- Step 2: We distinguish functions only up to multiplicative constants. E.g., $n/2$, n and $10n$ are indistinguishable,

These two decisions give us most of the robustness properties we desire, and are captured in the following language of asymptotics.

Where is Asymptopia? a far, far away land, where everything is BIG.

¶32. Eventuality. This is Step 1 in our search for invariance. Let $P : \mathbb{R} \to \{0, 1\}$ be a (partial) real predicate. We say P holds **eventually**, denoted "P (ev.)", if $P(x)$ holds for all x large enough. More precisely:

$$(\exists x_0)(\forall x)[x \geqslant x_0 \Rightarrow P(x)]. \tag{20}$$

Instead of "P (ev.)", we may also write

$$P(x) \text{ (ev. } x).$$

to explicitly show the role of the variable x. According to our rules for quantifying over partial predicates, "$(\forall x)$" in (20) really says "($\forall x$ such that $P(x) =\downarrow$)".

A typical example is when $P(x)$ is the predicate "$f(x) \leqslant g(x)$" defined by two complexity functions f, g. Then we say "$f \leqslant g$ (ev.)" if $f(x) \leqslant g(x)$ holds for all x large enough. More precisely,

$$(\exists x_0)(\forall x)[x \geqslant x_0 \Rightarrow f(x) \leqslant g(x)].$$

The "$(\forall x)$" in this statement really says "($\forall x$ such that $f(x) =\downarrow$ and $g(x) =\downarrow$)".

By not caring about the behavior of complexity function over some initial values, our complexity bounds becomes robust against the following **table-lookup trick**. If A is any algorithm, relative to to any given finite set S of inputs, we can modify A so that if $x \in S$, then the answer for x is obtained by a table lookup; otherwise, the answer is computed by running A on x. The modified algorithm A' might be much faster than A for all $x \in S$, but it will have the same "eventual" complexity as A. Thus, the complexity of A and A' are indistinguishable using our eventuality criterion.

¶33. Infinitely Often. The concept of eventuality is intimately connected with the concept of **infinitely often** ("i.o." for short). Given a real predicate $P(x)$, we say P holds **infinitely often**, written

$$P(x) \text{ (i.o. } x) \tag{21}$$

(or simply "P (i.o.)" when x is understood) if

$$(\forall x_0)(\exists x > x_0)[P(x)].$$

We have this logical equivalence:

$$\neg(P(x) \text{ (ev. } x)) \equiv (\neg P(x)) \text{ (i.o. } x) \tag{22}$$

In other words, there is a De Morgan-like law governing the pair of conditions (ev. x) and (i.o. x).

For instance, for complexity functions f and g, we say "$f \leqslant g$ (i.o.)" if for all x_0, there exists $x > x_0$ such that $f(x) =\downarrow$ and $g(x) =\downarrow$ and $f(x) \leqslant g(x)$. Note[13] that an "infinitely often" (i.o.) statement is equivalent to the negation of an "eventually" statement:

$$\neg[P(x) \text{ (ev. } x)] \equiv [\neg P(x)] \text{ (i.o. } x) \tag{23}$$

Most natural functions f in complexity satisfy some rather natural properties:

- f is eventually defined, $f(x) =\downarrow$ (ev.).
- f is eventually non-negative, $f \geqslant 0$ (ev.).

When these properties fail, our intuitions about complexity functions may go wrong.

¶34. Domination. We now take Step 2 towards invariance. We say f **dominates** g, written

$$f \succeq g,$$

if there exists $C > 0$ such that $C \cdot f \geqslant g$ (ev.). The symbol '$\succeq$' is intended to evoke the '$\geqslant$' connection. In particular, it should suggest the transitivity property: $f \succeq g$ and $g \succeq h$ implies $f \succeq h$. Of course, the reflexivity property holds: $f \succeq f$. We can also write "$g \preceq f$" instead of $f \succeq g$. If $f \succeq g$ and $g \succeq f$ then we write

$$f \asymp g.$$

Clearly $\asymp$ is an equivalence relation. The equivalence class of f is (essentially) the **Θ-order** of f; more on this below. If $f \succeq g$ but not $g \succeq f$ then we write

$$f \succ g$$

ans say that f **strictly dominates** g. E.g., $n^2 \succ n \succ 1 + \frac{1}{n}$. An interesting example is $\sin x \prec 1$. Note that $\sin x$ and $\cos x$ are not related by any of these relations. **Remark**: Collins [1] used the concept of domination and a notation very similar to ours:"$f \succeq g$" means $f \geqslant c \cdot g$ for some $c > 0$ (over the common domain of f and g). Thus, his notation does not have our eventuality condition. If $f \succeq g$ and $g \succeq f$, he writes "$f \sim g$".

3 *domination-type relations:* $\succeq, \succ, \asymp$

[13] A possibly confusion is this: if P (i.o.), then it is true that there are infinitely many values of x's such that $P(x)$ holds. However, this is not sufficient. We need the x's to increase without bound.

Thus the triplet of notations $\succeq, \succ, \asymp$ for real functions correspond to the binary relations $\geqslant, >, =$ for real numbers.

Domination provides "implementation platform independence" for our complexity results: it does not matter whether you implement a given algorithm in a high level program language like `Java` or in assembly language. The complexity of your algorithm in these implementations (if done correctly) will be dominated by each other (i.e., same Θ-order). This also insulates our complexity results against Moore's Law: over a limited time period, the timing of our algorithms keeps the same Θ-order. Of course, Moore's law cannot hold indefinitely because of physical limits, but the end is not in sight yet.

one form of Moore's law predicts that the "speed" of hardware will double every 18 months...

¶35. The Big-Oh Notation. We write

$$\mathcal{O}(f)$$

(and read **order of** f or **big-Oh of** f) to denote the set of all complexity functions g such that

$$0 \preceq g \preceq f.$$

the key asymptotic notation to know!

Note that each function in $\mathcal{O}(f)$ dominates 0, *i.e.*, is eventually non-negative. Thus, restricted to functions that are eventually non-negative, the big-Oh notation (viewed as a binary relation) is equivalent to domination.

big-Oh is almost the same as domination

> *We can unroll the big-Oh notation as follows: To prove $g = \mathcal{O}(f)$, you need to show some $C > 0$ and x_0 such that for all $x \geqslant x_0$, if $g(x) =\downarrow$ and $f(x) =\downarrow$ then $0 \leqslant g(x) \leqslant Cf(x)$. Remember your epsilon-delta argument in Calculus? Well, the Computer Science analogue is the C-x_0 argument.*

$\delta : \epsilon :: C : x_0$

E.g., The set $\mathcal{O}(1)$ comprises all functions f that are bounded and eventually non-negative. The function $1 + \frac{1}{n}$ is a member of $\mathcal{O}(1)$.

The simplest usage of this $\mathcal{O}$-notation is as follows: we write

$$g = \mathcal{O}(f)$$

(and read 'g **is big-Oh of** f' or 'g **is order of** f') to mean g is a member of the set $\mathcal{O}(f)$. The equality symbol '=' here is "uni-directional": $g = \mathcal{O}(f)$ does not mean the same thing as $\mathcal{O}(f) = g$. Below, we will see how to interpret the latter expression. The equality symbol in this context is called a **one-way equality**. Why not just use '$\in$' for the one-way equality? Main reason is convenience: if $f \in O(g)$, $g \in O(h)$, $h = O(\cdots)$, and we want to merge these statements, then we need to write $f \in O(g) \subseteq O(h) \subseteq \cdots$. This is awkward. With one-way equalities, we simply write $f = O(g) = O(h) = \cdots$. The equality symbol has a uni-directional flavor whereby we (repeatedly) transform a formula from an unknown form into a known form, separated by an equality symbol. For example, the following sequence of one-way equalities

$$f(n) = \sum_{i=1}^{n}(i + \frac{n}{i}) = \left(\sum_{i=1}^{n} i\right) + \left(\sum_{i=1}^{n} \frac{n}{i}\right) = \mathcal{O}(n^2) + \mathcal{O}(n \log n) = \mathcal{O}(n^2)$$

is a derivation to show f is at most quadratic.

¶* 36. Big-Oh Expressions. The expression '$\mathcal{O}(f(n))$' is an example of an $\mathcal{O}$-expression, which we now define. In any $\mathcal{O}$-expression, there is a **designated variable** which is the real variable that goes[14] to infinity. For instance, the $\mathcal{O}$-expression $\mathcal{O}(n^k)$ would be ambiguous were it not for the tacit convention that 'n' is normally the designated variable. Hence k is assumed to be constant. We shall define **$\mathcal{O}$-expressions** as follows:

(Basis) If f is the symbol for a function, then f is an $\mathcal{O}$-expression. If n is the designated variable for $\mathcal{O}$-expressions and c a real constant, then both 'n' and 'c' are also $\mathcal{O}$-expressions.

(Induction) If E, F are $\mathcal{O}$-expressions and f is a symbol denoting a complexity function then the following are $\mathcal{O}$-expressions:

$$\mathcal{O}(E), \quad f(E), \quad E+F, \quad EF, \quad -E, \quad 1/E, \quad E^F.$$

Each $\mathcal{O}$-expression E denotes a set $\widetilde{E}$ of partial real functions in the obvious manner: in the basis case, a function symbol f denotes the singleton set $\widetilde{f} = \{f\}$. Inductively, the expression $E + F$ (for instance) denotes the set $\widetilde{E+F}$ of all functions $f + g$ where $f \in \widetilde{E}$ and $g \in \widetilde{F}$. Similarly for

$$\widetilde{f(E)}, \quad \widetilde{EF}, \quad \widetilde{-E}, \quad \widetilde{E^F}.$$

The set $\widetilde{1/E}$ is defined as $\left\{1/g : g \in \widetilde{E} \;\&\; 0 \preceq g\right\}$. Finally,

$$\widetilde{\mathcal{O}(E)} = \left\{f : (\exists g \in \widetilde{E})[0 \preceq f \preceq g]\right\}.$$

Examples of $\mathcal{O}$-expressions:

$$2^n - \mathcal{O}(n^2 \log n), \qquad n^{n+\mathcal{O}(\log n)}, \qquad f(1+\mathcal{O}(1/n)) - g(n).$$

If E is an $\mathcal{O}$-expression, we say that E dominates 0, written $E \succeq 0$, if each $f \in \widetilde{E}$ dominates 0. Not all $\mathcal{O}$-expressions dominate 0. E.g., $-\mathcal{O}(E)$ does not dominate 0. But $2^{-\mathcal{O}(E)}$ does. If E, F are two $\mathcal{O}$-expressions, we may write

$$E = F$$

to denote $\widetilde{E} \subseteq \widetilde{F}$, *i.e.*, the "equality symbol" stands for set inclusion! This generalizes our earlier "$f = \mathcal{O}(g)$" interpretation. Some examples of this usage:

$$\mathcal{O}(n^2) - 5^{\mathcal{O}(\log n)} = \mathcal{O}(n^{\log n}), \qquad n + (\log n)\mathcal{O}(\sqrt{n}) = n^{\log\log n}, \qquad 2^n = \mathcal{O}(1)^{n-\mathcal{O}(1)}.$$

An ambiguity arises from the fact that if $\mathcal{O}$ does not occur in an $\mathcal{O}$-expression, it is indistinguishable from an ordinary expression. We must be explicit about our intention, or else rely on the context in such cases. Normally, at least one side of the one-sided equation '$E = F$' contains an occurrence of '$\mathcal{O}$', in which case, the other side is automatically assumed to be an $\mathcal{O}$-expression. Some common $\mathcal{O}$-expressions are:

- $\mathcal{O}(0)$, the eventually zero functions.
- $\mathcal{O}(1)$, the bounded functions.
- $1 \pm \mathcal{O}(1/n)$, a set of functions that tends to $1^{\pm}$.

[14] More generally, we can consider x approaching some other limit, such as 0.

- $\mathcal{O}(n)$, the linearly bounded functions.
- $n^{\mathcal{O}(1)}$, the functions bounded by polynomials.
- $\mathcal{O}(1)^n$ or $2^{\mathcal{O}(n)}$, the functions bounded by simple exponentials.
- $\mathcal{O}(\log n)$, the functions bounded by some multiple of the logarithm.

¶* 37. Extensions of Big-Oh Notations. We note some simple extensions of the $\mathcal{O}$-notation:
(1) **Inequality interpretation:** For $\mathcal{O}$-expressions E, F, we may write $E \neq F$ to mean that the set of functions denoted by E is not contained in the set denoted by F. For instance, $f(n) \neq \mathcal{O}(n^2)$ means that for all $C > 0$, there are infinitely many n such that $f(n) > Cn^2$.
(2) **Subscripting convention:** We can subscript the big-Oh's in an $\mathcal{O}$-expression. For example,

$$\mathcal{O}_A(n), \quad \mathcal{O}_1(n^2) + \mathcal{O}_2(n \log n). \tag{24}$$

The intent is that each subscript (A, 1, 2) picks out a specific but anonymous function in (the set denoted by) the unsubscripted $\mathcal{O}$-notation. Furthermore, within a given context, two occurrences of an identically subscripted $\mathcal{O}$-notation are meant to refer to the same function. For subscripted expressions, it now makes sense to use inequalities, as in "$f \geqslant \mathcal{O}_A(g)$" or "$f \leqslant \mathcal{O}_1(g)$".

For instance, if A is a linear time algorithm, we may say that "A runs in time $\mathcal{O}_A(n)$" to indicate that the choice of the function $\mathcal{O}_A(n)$ depends on A. Further, all occurrences of "$\mathcal{O}_A(n)$" in the same discussion will refer to the same anonymous function. Again, we may write

$$n2^k = \mathcal{O}_k(n), \quad n2^k = \mathcal{O}_n(2^k)$$

depending on one's viewpoint. Especially useful is the ability to do "in-line calculations". As an example, we may write

$$g(n) = \mathcal{O}_1(n \log n) = \mathcal{O}_2(n^2)$$

where, it should be noted, the equalities here are true equalities of functions.

(3) Another possible extension is to multivariate real functions. For instance consider the notation "$f(x, y) = \mathcal{O}(g(x, y))$" where we view both x and y are designated variables. I.e., there exist constants $C > 0, x_0, y_0$ such that for all $x > x_0, y > y_0$, $f(x, y) \leqslant Cg(x, y)$. In practice, such an extension is seldom needed.

¶38. Other Asymptotic Notations: small-oh and super-domination So far, we have define a pair of closely related concepts: a *binary relation* called domination $f \succeq g$, and an *order notation* called Big-Oh $g = O(f)$. The difference between the two concepts is that the order notation has additional requirements, namely $g = O(f)$ must satisfy $g \succeq 0$. We now introduce four more pairs of such asymptotic notations. This is given by the second and last columns of following table:

Name of $\odot$	Notation $g = \odot(f)$	Informal Meaning	Definition of order $g = \odot(f)$	Analogous In-fix Notation
big-Oh	$g = O(f)$	$g \leqslant f$	$(\exists C > 0)[C \cdot f \geqslant g \geqslant 0(\text{ev.})]$	$f \succeq g$ (dominates)
big-Omega	$g = \Omega(f)$	$g \geqslant f$	$(\exists C > 0)[C \cdot g \geqslant f \geqslant 0(\text{ev.})]$	$f \preceq g$ (dominated by)
Theta	$g = \Theta(f)$	$g = f$	$(\exists C > 1)[C^2 \cdot f \geqslant C \cdot g \geqslant f \geqslant 0(\text{ev.})]$	$g \asymp f$ (co-dominates)
small-oh	$g = o(f)$	$g \ll f$	$(\forall C > 0)[C \cdot f > g \geqslant 0(\text{ev.})]$	$f \gg g$ (super-dominates)
small-omega	$g = \omega(f)$	$g \gg f$	$(\forall C > 0)[C \cdot g > f \geqslant 0(\text{ev.})]$	$f \ll g$ (super-dominated by)

The four new order notations (analogous to $O(f)$) are

$$\text{big-Omega } \Omega(f); \text{ Theta } \Theta(f); \text{ small-oh } o(f); \text{ small-omega } \omega(f).$$

Their formal definitions are given by Column 4 of the table. The first 3 rows of the table (big-Oh, big-Omega, Theta) are easily understood in terms of the domination relation, $f \succeq g$. But the last 2 rows require a binary relation called "super-domination". We say f **super-dominates** g, written $f \gg g$ if

$$(\forall C > 0)[C \cdot f > g \text{ (ev.)}] \tag{25}$$

Note that we use $>$ instead of $\geqslant$ in (25). In our definition of domination $f \succ g$, we use $\geqslant$. In most applications, this subtlety change is irrelevant. However, our definition rules out "$0 \gg 0$".

$\gg$ is like the informal concept '$\gg$'

The corresponding set notations are called **small-oh/small-omega**, analogous to big-Oh/big-Omega. By definition, $f \gg g$ iff $g \ll f$. Note that $f \gg g$ implies $f \succ g$.

Remark: The literature often define "$g = o(f)$" or "$f \gg g$" by using limits, i.e., $g(x)/f(x) \to 0$ as $x \to \infty$. In the spirit of the non-calculus or "elementary approach", we simply avoid discussion of limits.

> CONVENTION: We observe that for any of the five choices of $\odot$, the set $\odot(f)$ is non-empty iff $f \succeq 0$. Henceforth, we adopt this convention: *whenever we write the expression "$\odot(f)$", it is assumed to be non-empty, i.e., $f \succeq 0$.*

We now unpack these definitions in a leisurely manner.

Big-Omega notation: $\Omega(f)$ is the set of all complexity functions g such that for some constant $C > 0$,

$$C \cdot g \geqslant f \geqslant 0 \text{ (ev.)}.$$

Of course, this can be compactly written as $g \succeq f \succeq 0$. Clearly, big-Omega is just the reverse of the big-Oh relation: $g = \Omega(f)$ iff $f = \mathcal{O}(g)$.

Theta notation: $\Theta(f)$ is the intersection of the sets $\mathcal{O}(f)$ and $\Omega(f)$. So g is in $\Theta(f)$ iff $g \asymp f$.

Small-oh notation: $o(f)$ is the set of all complexity functions g such that for all $C > 0$,

so C can be arbitrarily small!

$$C \cdot f > g \geqslant 0 \text{ (ev.)}.$$

As usual, we write $g = o(f)$ to mean $g \in o(f)$. For instance, with $f(x) = 1$ and $g(x) = 1/x$, we conclude that $1/x = o(1)$. Also, we have the relation $o(f) \subseteq \mathcal{O}(f)$.

Small-omega notation: $\omega(f)$ is the set of all functions g such that for all $C > 0$,

$$C \cdot g > f \geqslant 0 \text{ (ev.)}.$$

Clearly $\omega(f) \subseteq \Omega(f)$.

For each of these notations, we can again define the $\odot$-expressions ($\odot \in \{\Omega, \Theta, o, \omega\}$), use the one-way equality instead of set-membership or set-inclusion, and employ the subscripting convention. Thus, we write "$g = \Omega(f)$" instead of saying "g is in $\Omega(f)$". We call the set $\odot(f)$ the **$\odot$-order** of f. Here are some immediate relationships among these notations:

- $f = \mathcal{O}(g)$ iff $g = \Omega(f)$.
- $f = \Theta(g)$ iff $f = \mathcal{O}(g)$ and $f = \Omega(g)$.
- $f = \mathcal{O}(f)$ and $\mathcal{O}(\mathcal{O}(f)) = \mathcal{O}(f)$.
- $f + o(f) = \Theta(f)$.
- $o(f) \subseteq \mathcal{O}(f)$.
- $g = \omega(f)$ iff $f = o(g)$.

¶39. Varieties of Lower Bounds. varieties The relation between upper and lower bounds between two real numbers is straightforward. But this relation is more subtle for a pair of complexity functions: what do we mean when we say "g is a lower bound on f"? We want to explore this carefully because lower bounds concepts are often misused in the literature. In the following, it is simpler if we assume that $f, g \geqslant 1$ (ev.).

- The simplest way is to say that "g is a lower bound on f" is $f = \Omega(g)$. This translates into $f = \Omega(g)$
$$(\exists C > 0)(\exists n_0)(\forall n > n_0)[Cf(n) > g(n)]. \tag{26}$$
- But we could also negate the upper bound statement "$f = \mathcal{O}(g)$". Thus the statement $f \neq \mathcal{O}(g)$ gives another kind of lower bound on f: $f \neq \mathcal{O}(g)$
$$(\forall C > 0)(\forall n_0)(\exists n > n_0)[Cf(n) > g(n)]. \tag{27}$$
- Using the small-omega and small-oh notations, we have two similar ways to state lower bounds. Thus $f = \omega(g)$ translates into $f = \omega(g)$
$$(\forall C > 0)(\exists n_0)(\forall n > n_0)[f(n) > Cg(n)]. \tag{28}$$
- And finally $f \neq o(g)$ translates into $f \neq o(g)$
$$(\exists C > 0)(\forall n_0)(\exists n > n_0)[f(n) > Cg(n)]. \tag{29}$$

Notice that the matrix '$[f(n) > Cg(n)]$' is common to all four lower bound statements (26)–(29). Of these, two are direct application of our notations ($= \Omega(g)$ and $\omega(g)$) but two are *negations* of our notations ($\neq \mathcal{O}(g)$ and $\neq o(g)$). It can be seen from the above translations that the four lower bounds are related via these four implications:

$$\begin{array}{ccccc} & & f = \Omega(g) & & \\ & \nearrow & & \searrow & \\ f = \omega(g) & & & & f \neq o(g) \\ & \searrow & & \nearrow & \\ & & f \neq O(g) & & \end{array} \tag{30}$$

Likewise, we could introduce four ways of stating upper bounds.

Let us see how these notations are used in practice. For example, let us prove that for all $k < k'$,
$$n^{k'} \neq \mathcal{O}(n^k).$$
Suppose $n^{k'} = \mathcal{O}(n^k)$. Then there is a $C > 0$ such that $n^{k'} \leqslant Cn^k$ (ev.). That means $n^{k'-k} \leqslant C$ (ev.). This is a contradiction because n^ε is unbounded for any $\varepsilon > 0$.

¶* 40. What if Θ-order is too coarse? Despite our general aim of looking only at Θ-order of complexity functions, there is often need to refine this. One concept that is often used in the mathematical literature is this: write

$$f \sim g \tag{31}$$

if $f = g \pm o(g)$ or $f(x) = g(x)[1 \pm o(1)]$. This says that f and g approximate each other with relative error of $o(1)$.

so $n \sim n + \lg n$ but $n \not\sim 2n$.

Let us illustrate the need for this notation using the **unbounded search problem**: given an unknown real number x^*, we want to an interval $[a, b]$ containing x^* such that $b - a \leqslant 1$. When we have found such an interval, we say that x^* has been "located". We assume an **oracle for** x^*. We can pose queries of the form "is $x^* \leqslant x$?" (for any chosen $x \in \mathbb{R}$). The oracle will either answer yes or no. Based on this answer, we can pose another query, etc. How many queries must we ask before we can locate x^*? It is relatively to solve the unbounded search problem using

this is like the comparison tree model, but the tree is now infinite.

$$T_1(x^*) = 2\lg(|x^*|) + O(1) \tag{32}$$

queries. But we can also solve it with

$$T_2(x^*) = \lg(|x^*|) + 2\lg\lg(|x^*|) + O(1) \tag{33}$$

queries, or even better,

$$T_3(x^*) = \lg(|x^*|) + \lg\lg(|x^*|) + \lg\lg\lg(|x^*|) + O(1). \tag{34}$$

The problem of unbounded search will be taken up in Chapter 12. For now, we notice that $T_1 \not\prec T_2$, but $T_2 \sim T_3$. In order to distinguish T_3 from T_2, we propose the following definition: let

$$f \stackrel{+}{\preceq} g \iff 2^f \preceq 2^g \tag{35}$$

Similarly, $f \stackrel{\pm}{\simeq} g$ iff $f \stackrel{+}{\preceq} g$ and $g \stackrel{+}{\preceq} f$. Finally, $f \stackrel{+}{\prec} g$ iff $f \stackrel{+}{\preceq} g$ but not $g \stackrel{+}{\preceq} f$. Now, we see that

$$T_3 \stackrel{+}{\prec} T_2.$$

Call these $\stackrel{+}{\preceq}$, $\stackrel{+}{\prec}$, $\stackrel{\pm}{\simeq}$ the **+-domination** concepts. Intuitively, they provide approximations up to additive constants, in contrast to the standard domination concepts ($\preceq$, $\prec$, $\simeq$) which are approximations up to multiplicative constants.

¶* 41. Discussion of asymptotic notations. There is some debate over the best way to define the asymptotic concepts in computer science. So it is not surprising that there is considerable divergence on the details in the literature. Here we note just two alternatives:

be warned!

- Perhaps the most common definition follows Knuth [6, p. 104] who defines "$g = \mathcal{O}(f)$" to mean there is some $C > 0$ such that $|f(x)|$ dominates $C|g(x)|$. Using this definition, both $\mathcal{O}(-f)$ and $-\mathcal{O}(f)$ would mean the same thing as $\mathcal{O}(f)$. But our definition allows us to distinguish[15] between $1 + \mathcal{O}(1/n)$ and $1 - \mathcal{O}(1/n)$. Note that $g = 1 - \mathcal{O}(f)$ amounts to $1 - Cf \leqslant f \leqslant 1$ (ev.). When an big-Oh expression appears in negated form as in $-\mathcal{O}(1/n)$, it is really a lower bound

[15] On the other hand, there is no easy way to recover Knuth's definition using our definitions. It may be useful to retain Knuth's definition by introducing a special notation "$|\mathcal{O}|(f(n))$", etc.

- Again, we could have defined "$\mathcal{O}(f)$" more simply, as comprising those g such that $g \preceq f$. That is, we omit the requirement $0 \preceq g$ from our original definition. This alternative definition is attractive for its simplicity. But the drawback of this simplified "$\mathcal{O}(f)$" is that it contains arbitrarily negative functions. The expression $1 - \mathcal{O}(1/n)$ is useful as an upper and lower bound under our official notation. But with the simplified definition, the expression "$1 - \mathcal{O}(1/n)$" has no value as an upper bound. Our official definition opted for something that is intermediate between this simplified version and Knuth's.

We are following Cormen et al [2] in restricting the elements of $\mathcal{O}(f)$ to complexity functions that dominate 0. This approach has its own burden: thus whenever we say "$g = \mathcal{O}(f)$", we have to check that g dominates 0 (cf. exercise 1 below). In practice, this requirement is not much of a burden, and is silently passed over.

A common abuse is to use big-Oh notations in conjunction with the inequality symbol ($\leqslant$). It is very tempting to write "$f(n) \leqslant \mathcal{O}(g)$" instead of the correct "$f(n) = \mathcal{O}(g)$". At best, writing "$\leqslant$" is redundant. The problem is that, once this redundancy is admitted, one may in the course of a long derivation eventually write "$f(n) \geqslant \mathcal{O}(E)$" which is not meaningful. Hence we regard any use of "$\leqslant$" or "$\geqslant$" in connection with $\mathcal{O}$-notations as illegitimate (but it is legitimate again under the subscripting convention (24)).

Perhaps most confusion (and abuse) in the literature arises from variant definitions of the Ω-notation. For instance, one may have only shown a lower bound of the form $f \neq \mathcal{O}(g)$ or $f \neq o(g)$ result, but this is viewed as a proof of $f = \Omega(g)$ or $g = \omega(g)$. We see from (30) that these are quite different.

Evidently, these asymptotic notations can be intermixed. E.g., $o(n^{\mathcal{O}(\log n)} - \Omega(n)$. However, they can be tricky to understand and use, and there seems to be little need for them. Another generalization with some applications are multivariate complexity functions such as $f(x, y)$. They do arise in discussing tradeoffs between two or more computational resources such as space-time, area-time, etc. In recently years, the study of "parametrized complexity" has given example of multivariate complexity functions where some of the size variables controls the "parameters" of the problem.

Exercises

Exercise 8.1: What is the relation between $A \equiv (P(x) \wedge Q(x))$ (i.o. x) and $B \equiv (P(x)$ (i.o. $x)) \wedge (Q(x)$ (i.o. $x))$ ♢

Exercise 8.2: We defined the relation $f \gg g$ as follows:

$$(\forall C > 0)[C \cdot f > g \text{ (ev.)}].$$

Suppose we modify it to be

$$(\forall C > 0)[C \cdot f \geqslant g \text{ (ev.)}].$$

Give an example of one relation $f \gg g$ that holds under the new definition, but not in our official definition. ♢

Exercise 8.3: (i) (Eventually)
Show that $10x \leqslant 0.1x^2 - 100$ (ev.)

(ii) (Unboundedness of Harmonic numbers)
For any constant K, show that $H_n > K$ for integers n large enough.
Here $H_n := 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$ is called the n-th Harmonic number. For instance $H_3 = 11/6$. We may express the result as "$H_n > K$ (ev. n)" even though n are restricted to positive integers.
HINT: let K a positive integer n. Consider H_{2^n} and divide the terms of H_{2^n} into n groups where each group sums to at least $1/2$.

(iii) (Strict domination of polynomials by exponentials)
$e^x \succ x^k$ for any positive k.
HINT: Use the definition $e^x := \sum_{i \geqslant 0} x^i/i!$. See Appendix, Chap.II for more information.

♢

Exercise 8.4: Below are two groups of 4 conditions each. Give the Hasse diagram of implications among conditions in each group. To show that your Hasse diagram has all possible relations, illustrated by functions that violate implications not in your Hasse diagram.
(A)

(A-i) $f > 0$ (ev.)

(A-ii) $f \geq 0$

(A-iii) $f > 0$

(A-iv) $f \gg 0$

(B) Give the Hasse diagram of implications among these conditions.

(B-i) $f > 1$ (ev.)

(B-ii) $f \geq 1$

(B-iii) $f > 1$

(B-iv) $f \gg 1$

♢

Exercise 8.5: (a) Suppose that for all $C > 0$, we have $f > Cg$ infinitely often (i.o.). Please express this using our asymptotic notations (like dominance, etc).
(b) Please restate the condition $f \not\ll g$ (f is not super-dominated by g) using the "infinitely often" terminology.
(c) Show two functions f, g such that $f \prec g$ but $f \not\ll g$. ♢

Exercise 8.6: Our asymptotic notations falls under two groups: O, Ω, Θ and o, ω. In the first group, we have $\Theta(f) = O(f) \cap \Omega(f)$. This suggests the "small-theta" analogue for the second group, "$\theta(f) = o(f) \cap \omega(f)$". Why was this not done? ♢

Exercise 8.7: Let $P : D \to \{0, 1\}$ be a partial predicate over some domain D. When we do quantification, $\forall x$ and $\exists x$ it is assumed that x range over D. Show that the following equivalences (called "de Morgan's laws for quantifiers") hold:
(a) $\neg(\forall x)P(x)$ is equivalent to $(\exists x)\neg P(x)$
(b) $\neg(\exists x)P(x)$ is equivalent to $(\forall x)\neg P(x)$ ♢

Exercise 8.8: To do this problem, we recall some common mathematical expressions:
(i) "f is unbounded" means that for any $C > 0$, there exists x such that $f(x) > C$.
(ii) "$f > g$ (i.o.)" (i.o. = infinitely often) means that there are arbitrarily large values of x where $f(x) > g(x)$ holds.
(iii) "f is bounded away from 0" means there exists $\epsilon > 0$ such that for all x, $f(x) \geqslant \epsilon$.

(a) Condition is "$f \gg 1$". Show an f that is unbounded but does not satisfy this condition.

(b) Condition is "$f \not\preceq 1$" (i.e., "It is not the case that $f \preceq 1$"). Give an English expression for this condition.

(c) Condition is "$f \succ 1$". Give an English expression for this condition.

(d) Clearly, Condition (a) implies Condition (b). Give a counter example for the converse.

♢

Exercise 8.9: Assume $f(n) \geqslant 1$ (ev.).
(a) Show that $f(n) = n^{\mathcal{O}(1)}$ iff there exists $k > 0$ such that $f(n) = \mathcal{O}(n^k)$. This is mainly an exercise in unraveling our notations!
(b) Show a counter example to (a) in case $f(n) \geqslant 1$ (ev.) is false. ♢

Exercise 8.10: Prove or disprove: $f = \mathcal{O}(1)^n$ iff $f = 2^{\mathcal{O}(n)}$. ♢

Exercise 8.11: If $P_i : D \to \{0, 1\}$ are partial predicates ($i = 0, 1$) over some domain D, then so are $\neg P_i$, $P_0 \vee P_1$ and $P_0 \wedge P_1$ where we use the rule that $\neg P_0(x), P_0(x) \vee P_1(x), P_0(x) \wedge P_1(x)$ are all undefined when $P_0(x) = \uparrow$. Show that $\neg(\forall x)P(x)$ is equivalent to $(\exists x)\neg P(x)$ and $\neg(\exists x)P(x)$ is equivalent to $(\forall x)\neg P(x)$. NOTE: these are called De Morgan's law for quantifiers, which is well-known when P is a total predicate. ♢

Exercise 8.12: Unravel the meaning of the $\mathcal{O}$-expression: $1 - \mathcal{O}(1/n) + \mathcal{O}(1/n^2) - \mathcal{O}(1/n^3)$. Does the $\mathcal{O}$-expression have any meaning if we extend this into an infinite expression with alternating signs? ♢

Exercise 8.13: For basic properties of the logarithm and exponential functions, refer to the Appendix in Chapter II. In the following, n is the designated variable, but c, k are constants. To prove a relation, you must explicitly specify the numerical constants (e.g., $n_0 = 3$ or $C = 2.5$) implicit in the asymptotic notations. *Try to use only elementary arguments.* Please prove the following:
(a) $(n + c)^k = \Theta(n^k)$. Note that c, k can be negative.
(b) $\log(n!) = \Theta(n \log n)$.
(c) $n! = o(n^n)$.
(d) $\lceil \log n \rceil ! = \Omega(n^k)$ for any $k > 0$.
(e) $\lceil \log \log n \rceil ! \leqslant n$ (ev.). ♢

Exercise 8.14: Show that $\ln(n!) = n \ln n + \Theta(n)$. You must not use Stirling's formula, but use the fact that the Harmonic numbers $H_n = \ln n + \Theta(1)$. ♢

Exercise 8.15: True or false: please provide a counter-example when false, and a proof when true. The base b of logarithms is arbitrary but fixed, and $b > 1$. Assume the functions f, g are arbitrary. Do not assume that f and g are $\geqslant 0$ eventually – the follow-up question next make assume additional properties of f, g.

(a) $f = \mathcal{O}(g)$ implies $g = \mathcal{O}(f)$.

(b) $\max\{f, g\} = \Theta(f + g)$.

(c) If $g > 1$ and $f = \mathcal{O}(g)$ then $\ln f = \mathcal{O}(\ln g)$.

(d) $f = \mathcal{O}(g)$ implies $f \circ \log = \mathcal{O}(g \circ \log)$. Assume that $g \circ \log$ and $f \circ \log$ are complexity functions.

(e) $f = \mathcal{O}(g)$ implies $2^f = \mathcal{O}(2^g)$.

(f) $f = o(g)$ implies $2^f = \mathcal{O}(2^g)$.

(g) $f = \mathcal{O}(f^2)$.

(h) $f(n) = \Theta(f(n/2))$.

♢

Exercise 8.16: Re-solve the previous exercise, assuming that $f, g \geqslant 2$ (ev.) and $f + g =\downarrow$ (ev.). ♢

Exercise 8.17: Let $f(x) = \sin x$ and $g(x) = 1$.
(i) Prove $f \preceq g$ or its negation.
(ii) Prove $g \preceq f$ or its negation.
HINT: To prove that $f \not\preceq g$, you need to show that for *all* choices of $C > 0$ and $x_0 > 0$, some relationship between f and g fails. ♢

Exercise 8.18: Suppose $T_A(n)$ is the running time of an algorithm A. We consider some notions of lower bounds on $T_A(n)$:
(a) Suppose you have constructed an infinite sequence of inputs $I_1, I_2, \ldots$ of sizes $n_1 < n_2 < \cdots$ such that A on I_i takes time more than $f(n_i)$. How can you express this lower bound result using our asymptotic notations?
(b) In the spirit of (a), what would it take to prove a lower bound of the form $T_A(n) \neq \mathcal{O}(f(n))$? What must you show about of your constructed inputs $I_1, I_2, \ldots$.
(c) What does it take to prove a lower bound of the form $T_A(n) = \Omega(f(n))$?
(d) Can you think of other ways to express lower bounds? ♢

Exercise 8.19: Using our order notations $O(f), \Omega(f), o(f), \omega(f)$ to answer this question:
(i) State four ways to say that "f is an upper bound on g".
(ii) Give any logical relations among the four ways in (i), assuming that $g \geqslant 0$ (ev.). If there is no relation, give a counter example. ♢

Exercise 8.20: Show some examples where you might want to use "mixed" asymptotic expressions. ♢

Exercise 8.21: Suppose $P(x, y)$ is a partial predicate, and $Q(y)$ is $(\forall x)P(x, y)$. Using our definitions, $Q(y)$ is now a total predicate. Should we modify our treatment of quantifiers to allow $Q(y)$ to be a partial predicates? ♢

Exercise 8.22: Discuss the meaning of the expressions $n - \mathcal{O}(\log n)$ and $n + \mathcal{O}(\log n)$ under (1) our definition, (2) Knuth's definition and (3) the "simplified definition" in the discussion. ♢

End Exercises

§9. Conclusion: Two Dictums of Algorithmics

To conclude this overview of algorithmics, we state two principles in algorithmics. They justify many of our procedures and motivate some of the fundamental questions we ask.

¶42. (D1) *Complexity functions are determined only up to Θ-order.* This recalls our motivation for introducing asymptotic notations, namely, concern for robust complexity results. For instance, we might prove a theorem that the running time $T(n)$ of an algorithm is "linear time", $T(n) = \Theta(n)$. Then simple and local modifications to the algorithm, or reasonable implementations on different platforms, should not affect the validity of this theorem.

There are many important caveats. We conclude from this dictum that it is important to design new algorithms with better Θ-complexity (such algorithms attain new "records" in the race towards optimality). While this attitude is good in itself, the converse attitude can be counter productive: we must not infer that only algorithms that achieve new records are important. Often, an asymptotically superior algorithm may be much slower than a slower algorithm when run on inputs of realistic sizes. For some problems, we might be interested in the constant multiplicative factors hidden by the Θ-notation. We also know that our ability to capture the simultaneous complexity of more than one computational resource is very limited. Finally, there are non-complexity issues that may matter. Simplicity of an algorithm is always appealing, in a non-quantifiable way. Ease-of-implementation might trump a purely complexity-based criterion. In short, we need a holistic view of algorithmics.

¶43. (D2) *Problems with complexity that are polynomial-bounded are feasible. Moreover, there is an unbridgeable gap between polynomial-bounded problems and those that are not polynomial-bounded.* This principle goes back to A. Cobham and J. Edmonds in the late sixties and relates to the P versus NP question. Hence, the first question we ask concerning any problem is whether it is polynomially-bounded. The answer may depend on the particular complexity model. E.g., a problem may be polynomial-bounded in space-resource but not in time-resource, although at this moment it is unknown if this possibility can arise. Of course, polynomial-bounded complexity $T(n) = n^c$ is not practical except for small c (typically less than 6). In many applications, even $c = 2$ is not practical. So the "practically feasible class" is a rather small slice of P.

Despite caveats, the dictums (D1) and (D2) turn out to be extremely useful. The landscape of computational problems is thereby simplified and made "understandable". The quest for asymptotically good algorithms helps us understand the nature of the problem. Often, after a complicated but asymptotically good algorithm has been discovered, we find ways to achieve the same asymptotic result in a simpler (practical) way.

The Babbage principle: the epigraph at the beginning of this Chapter may be interpreted as the "garbage in, garbage out" (GIGO) principle. In many areas of computation, it is not easy recognize when the input is meaningless. E.g., suppose we want to reconstruct a closed surface from a set of 3D points. If the points are not sufficiently closely sampled, there is no unique surface and it could be regarded as garbage for the algorithm. In this case, the only sensible output is a flag "GIGO". But the ability to detect a GIGO input is far from trivial. Surface reconstruction algorithms that always output a surface is clearly producing nonsense in such cases. This can fool the unwary user and is possibly dangerous.

Exercises

Exercise 9.1: How Big is 2^{64} in time (seconds) or space (grains of rice)?

Dictum (D2) says that algorithms with "exponential complexity" are not practical. This exercise underscores will underscore this message. In this question, you must use pen-paper estimates for orders of magnitude.

On Pen-Paper Estimates: In the spirit of our book's emphasis on "hand simulations" of algorithms, we also want you to do simple order-of-magnitude calculations. We want estimate numbers up to their "order of magnitude". I will define the **order of magnitude** of a number $T > 0$ as the integer power of 10 that is closest to T. Thus, $\lfloor \log_{10} T \rceil$ is the order of magnitude of T. E.g., the order of magnitude of the number 2^{20} is 6, since $2^{20} \sim 10^6$. Note that it is actually easier and *more accurate* to exploit the "Computer Science estimates" such as $2^{10} \simeq 1000 = 1K$ (kilo) and $2^{20} \simeq 10^6 = 1M$ (mega). After you have your Computer Science estimates, it is trivial to convert to the standard order-of-magnitude.

(a) Recall the "Tower of Hanoi Puzzle" on 64 disks. You may know from a programming course that we can solve this recursively in $2^{64} - 1$ moves (and this is optimal). **Question:** *If each move takes 1 second, estimate how many Big Bangs would it take to solve this puzzle?* Assume that one Big Bang is 14 billion years.

NOTE: Using a calculator, you can determine the exact answer to be about 41 Big Bangs (or 41.78 to be precise).

(b) The above is about time complexity. Let us consider the space complexity analogue: "Grains on a Chess Board" tells the story of an Indian king who lost a game of chess to a sage. The sage asked for his payment in rice: put one grain of rice in the first square of the chess board, put two grains in the second square, etc. Keep doubling the number of grains in successive squares until the 64th square. **Question:** *Are there enough atoms in the universe for the king to pay his debt?* Assume that a grain of rice has 10^{23} atoms, and the universe has about 10^{82} atoms. The answer is either Yes or No, but you must justify your answer.

♢

Exercise 9.2: How large a problem can we solve?

As usual, please use pen-paper estimates.

(a) (Time Limitations)
In general, the n-disk Tower of Hanoi Puzzle can be solved optimally in $2^n - 1$ moves. Suppose you have a super fast robot that can make a million moves per second. **Question:** *What is the largest n for which such a robot can solve the Puzzle within one year?*

(b) (Space Limitations)
What is the smallest $n \times n$ chessboard such that there is not enough grains of rice in the universe to repay the Indian sage?

♢

Exercise 9.3: True Upper and Lower Bound Estimates
In the above question, we estimated 2^{64} seconds to the nearest number of Big Bangs.

(a) Please give an upper bound on 2^{64} seconds.

(b) Please give a lower bound on 2^{64} seconds.

♢

Exercise 9.4: Monkeys at a Type-Writer
It is often said (in support of Evolution) that a monkey, given enough time, can reproduce the entire corpus of Shakespeare, the greatest writer of the English language. Estimate the time for this monkey to type the following sentence from the play *As you like it*:

All the world is a stage

The sentence has 23 characters and we do not distinguish small and capital letters. Assume a keyboard with 32 keys only, and the monkey can type continuously, at the rate of one million strokes per second. We can stop as soon as the monkey types the above sequence.

We further simplify the problem as follows: assume that we keep a counter $c \geqslant 0$, initialized to 0. If the last c characters is a prefix of the desired sentence, and the next character typed by the monkey is also correct, we increment c. Otherwise we reset $c \leftarrow 0$. Since the monkey is typing randomly, the question becomes: *what is the expected waiting time until the counter reaches $c = 23$ for the first time?*

♢

End Exercises

§10 APPENDIX A: General Notations

We gather some general notations used throughout this book. Use this as reference. If there is a notation you do not understand from elsewhere in the book, this is a first place to look.

Bookmark this appendix. Come back often!

¶A.0 Definitions.
We use the symbol $:=$ to indicate the definition of a term: we will write $X := \dots Y \dots$ when defining a term X in terms of $\dots Y \dots$ (known as the **definiens**). For example, we define the sign function as follows:

$$\texttt{sign}(x) := \begin{cases} 1 & \text{iff } x > 0 \\ 0 & \text{iff } x = 0 \\ -1 & \text{iff } x < 0 \end{cases}$$

Or, to define the special symbol for logarithm to base 2, we write $\lg x := \log_2 x$.

¶A.1 Numbers.
Denote the set of natural numbers[16] by $\mathbb{N} = \{0, 1, 2, \dots\}$, integers by $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$, rational numbers by $\mathbb{Q} = \{p/q : p, q \in \mathbb{Z}, q \neq 0\}$, the reals $\mathbb{R}$ and complex numbers $\mathbb{C}$. Thus we have

$$\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}.$$

The positive and non-negative reals are denoted $\mathbb{R}_{>0}$ and $\mathbb{R}_{\geqslant 0}$, respectively. Given $i, j \in \mathbb{N}$, let

$$[i..j] \quad \text{and} \quad [i..j)$$

denote the sets $\{i, i+1, \dots, j-1, j\}$ and $\{i, i+1, \dots, j-1\}$ (respectively). These sets are empty when $j < i$ (resp., $j \leqslant i$). If r is a real number, let its **ceiling** $\lceil r \rceil$ be the smallest integer greater than or equal to r. Similarly, its **floor** $\lfloor r \rfloor$ is the largest integer less than or equal to r. Clearly, $\lfloor r \rfloor \leqslant r \leqslant \lceil r \rceil$. For instance, $\lfloor 0.5 \rfloor = 0$, $\lfloor -0.5 \rfloor = -1$ and $\lceil -2.3 \rceil = -2$. The **fractional part** of a number r is $r - \lfloor r \rfloor$ and is often denoted $\{r\}$ (not to be confused with the singleton set, see below). The notation $\lceil r \rceil$ and $\lfloor r \rfloor$ is from the Computer Scientist Ken Iverson [6, p. 37].

For any positive integer m, let $\mathbb{Z}_m := \{0, 1, \dots, m-1\}$. This set forms a ring in which we can do addition, subtraction and multiplication modulo m. When m is prime, it is even a field in which we can do division by any non-zero element.

¶A.2 Sets and Multisets.
The **size** or **cardinality** of a set S is the number of elements in S and denoted $|S|$. The empty set is $\varnothing$. A set of size one (resp., two) is called a **singleton** (resp., **doublet**). The **disjoint union** of two sets is denoted $X \uplus Y$. This sometimes defined as

$$\{(x, 0) : x \in X\} \cup \{(y, 1) : y \in Y\}.$$

The intent to be able to uniquely attribute each element of $X \uplus Y$ to either X or Y. Of course, if $X \cap Y$ is empty, then there is no need for this artifice. We prefer to avoid this artifice and interpret "$X \uplus Y$" as simply $X \cup Y$ with the *side condition* that $X \cap Y = \varnothing$. In our definition, we can naturally say that if $z \in X \uplus Y$, then $z \in X$ or $z \in Y$, but not both. In particular, we may write $X = X_1 \uplus X_2 \uplus \cdots \uplus X_n$ to denote a **partition** of X into n subsets, i.e., the X_i's are pairwise disjoint and their union is X.

[16] Zero is considered natural here, although the ancients do not consider it so. The symbol $\mathbb{Z}$ comes from the German word 'Zahl' for numbers or 'zahlen' to count.

We often need to consider **multisets**. We assume the reader is familiar with the usual notion of sets, which we might call **standard sets** for emphasis. Multisets are sets whose elements need not be distinct. E.g., the multiset $S = \{a, a, b, c, c, c\}$ has 6 elements but only three of them are distinct, and it corresponds to the standard set $\underline{S} = \{a, b, c\}$. There are two copies of a and three copies of c in S. In general, for any multiset S, there is a standard set denoted $\underline{S}$, called the **underlying set** of S. Since we use the usual set braces to write multisets, the notation is ambiguous unless we explicitly say whether we are discussing sets or multisets. Alternatively, a multiset can be viewed as a function $\mu : S \to \mathbb{N}$ whose domain is a standard set S. Intuitively, $\mu(a)$ is the multiplicity of each $a \in S$. Of course, we could further extend multisets to weighted sets by letting $\mu : S \to \mathbb{R}_{\geqslant 0}$, etc.

¶A.3 Sets Constructions.

Given sets $X, X_1, X_2, \ldots$, we can define new sets using several standard operations:

1. Let 2^X denote the set of all subsets of X, called the **power set** of X. E.g., if $X = \{a, b\}$ then $2^X = \{\varnothing, \{a\}, \{b\}, \{a, b\}\}$.
2. If $n \in \mathbb{N}$ then an **n-set** refers to a set with cardinality n. A 1-set is also called a **singleton**. Let $\binom{X}{n}$ denote the set of n-subsets of X. Thus $\binom{X}{n} \subseteq 2^X$, and $\binom{X}{n}$ is empty if $n > |X|$. E.g., if $X = \{a, b, c\}$ then $\binom{X}{2} = \{\{a, b\}, \{b, c\}, \{c, a\}\}$.
3. The **Cartesian product** $X_1 \times \cdots \times X_n$ (also denoted $\prod_{i=1}^n X_i$) is the set of all n-tuples of the form $(x_1, \ldots, x_n)$ where $x_i \in X_i$. If $X_1 = \cdots = X_n$ then we simply write this as X^n.
4. A **sequence** in X is a finite list of the form $(x_1, x_2, \ldots, x_n)$ where each $x_i \in X$. The **length** of the sequence is $n \in \mathbb{N}$. The case $n = 0$ is the **empty sequence** $()$. Sometimes, sequences are also known[17] as **word** or **strings**. We usually denote a word $(x_1, x_2, \ldots, x_n)$ by juxta-position of its elements: $x_1 x_2 \cdots x_n$. Then the empty word is denoted ϵ. Let X^* denote the set of sequences (or words) over X. E.g., if $X = \{a\}$ then X^* is the infinite set $\{a, aa, aaa, \ldots\}$. We extend the notion of sequences to **infinite sequences**, $(x_1, x_2, \ldots)$ of length $n = \infty$. Let X^∞ denote the set of infinite sequences over X. Finally, let $X^\omega := X^* \cup X^\infty$.

¶A.4 Relations and Order.

An n-ary **relation** R on a set X is any subset of X^n: $R \subseteq X^n$. The most important case is when $n = 2$: R is then called a **binary relation**, and if $(a, b) \in R$, we like to write aRb and say "(a, b) is a **relationship** of R". We consider four properties that a binary relation R might have:

- [R]: **reflexive** if $(\forall a \in X)[aRa]$;
- [T]: **transitive** if $(\forall a, b, c \in X)[aRb$ and bRc implies $aRc]$;
- [S]: **symmetric** if $(\forall a, b \in X)[aRb$ implies $bRa]$;
- [A]: **anti-symmetric** if $(\forall a, b \in X)[aRb$ and bRa implies $a = b]$.

A **pre-order** R is a reflexive and transitive binary relation. For any $x \in X$, let $\overline{x} := \{y \in X : xRy, yRx\}$, called an **$R$-equivalence class**. The R-equivalence classes partitions X into disjoint subsets called **blocks**. We say $x, y \in X$ are **R-equivalent** if $\overline{x} = \overline{y}$. E.g.,

[17] The "word" terminology comes from formal language theory, where a "language" is a subset of X^*.

let $X = \mathbb{Z}$ and $xR_|y$ if x divides y (i.e., there exists $z \in \mathbb{Z}$ such that $xz = y$). Then $R_|$ is a pre-order on $\mathbb{Z}$. The $R_|$-equivalence classes are $\{n, -n\}$ for all $n \in \mathbb{Z}$. We often write $x|y$ instead of $xR_|y$.

A pre-order R that is symmetric is called an **equivalence relation**. In this case, we commonly write "$x \equiv y$" instead of xRy. A pre-order R that is anti-symmetric is called a **partial order**. In this case, we commonly write "$x \leqslant y$" instead of xRy. [18] Both $\equiv$ and $\leqslant$ are among the most fundamental binary relation of mathematics. The student may prove this lemma to practice the above definitions.

$\equiv$: [R], [S], [T].
$\leqslant$: [R], [A], [T].

Lemma 3 *Let $R \subseteq X^2$ be a pre-order.*
(i) The set $\overline{X} := \{\overline{x} : x \in X\}$ forms a partition of X. Recall $\overline{x}$ is an R-equivalence class of x.
(ii) The relation $\overline{R} \subseteq \overline{X}^2$ where $\overline{x}\overline{R}\overline{y}$ if xRy is a partial order on $\overline{X}$.
(iii) The relation R is a partial order iff $\overline{x} = \{x\}$ for all $x \in X$.
(iv) The relation R is an equivalence relation iff (for all $x \in X$)[xRy iff $\overline{x} = \overline{y}$].

A very useful concept in analysis of comparison based algorithms in the concept of a "Hasse Diagram". Fix a partial order $R \subseteq X^2$ on the set X. For $x, y \in X$, we write $x \leqslant y$ instead of xRy. If $x \leqslant y$ but $x \neq y$, we write $x < y$ (equivalently, $y > x$). We say a relationship $x > y$ is **essential** if there is no z such that $x > z > y$. We also say x **covers** y. The **Hasse diagram** of R is a directed graph $H(R)$ whose vertex set is X and for all $x, y \in X$, we have an edge x–y iff x covers y. For example, if R is a total order on X, then $H(R)$ is just a linear digraph.

after Helmut Hasse (1898-1979)

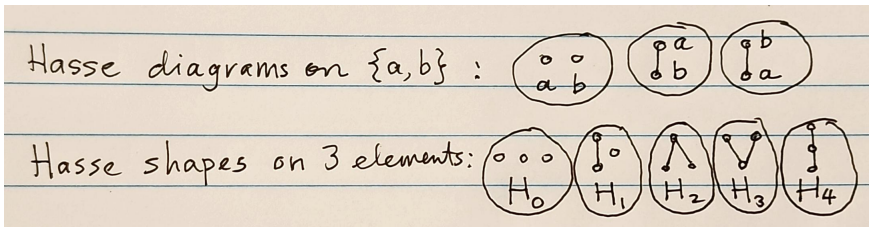


Figure 9: (i) Hasse diagrams on $X = \{a, b\}$; (ii) Hasse shapes of size 3

CONVENTION: In figures, we will draw the digraph $H(R)$ such that if x covers y, then x appears at a higher (horizontal) level than y. So the edge x–y is implicitly directed from higher to a lower level.
If the partial order R had been deduced from a comparison-based algorithm, and x covers y, then clearly the comparison $x : y$ must have been made by the algorithm (it could be deduced from the results of other comparisons). The comparison $x : y$ is then called **essential**. This concept is used to prove lower bounds on the comparison complexity.

In Figure 9(i), we show all the Hasse diagrams on $X = \{a, b\}$. There are only 3 of them. It would be a bit tedious to show all the 15 Hasse diagrams on $X = \{a, b, c\}$. Instead, we show a more abstract notion called **Hasse shapes** in Figure 9(ii). You can think of Hasse shapes as "unlabeled Hasse diagrams" There are (thankfully) only 5 of them.

[18] Mnemonic device: The 3 axioms for an equivalence relation on X are [R], [S], [T] (corresponding to 3 consecutive letters of the Latin alphabet. Furthermore, the 3 axioms involve (respectively) 1, 2, 3 elements of X. There is a similar mnemonic for a partial order, as the 3 axioms are [R], [A], [T].

Suppose H_i $(i = 1, 2)$ is a Hasse diagrams on X_i. We say H_1 and H_2 are equivalent if they are isomorphic as graphs. This means that, by renaming the elements in X_1, we can turn H_1 into H_2. Of course this implies $|X_1| = |X_2|$. Let $[H_1]$ denote the equivalence class of H_1. We call $[H_1]$ a **Hasse shape**. As suggested in the previous paragraph, we can[19] think of $[H_1]$ as the "unlabeled version" of H_1.

¶A.5 Functions.

If $f : X \to Y$ is a partial function, then write $f(x) = \uparrow$ if $f(x)$ is undefined and $f(x) = \downarrow$ otherwise. If for all x, $f(x) = \downarrow$, then f is a **total function**. Some authors use "$f : X \dashrightarrow Y$" to indicate partial functions, and reserve "$f : X \to Y$" for total functions. To avoid introducing a new notation, we assume $f : X \to Y$ is a total function by default, and say "$f : X \to Y$ is a partial function" when needed. Function composition will be denoted $f \circ g : X \to Z$ where $g : X \to Y$ and $f : Y \to Z$. Thus $(f \circ g)(x) = f(g(x))$. We need the special rule that when $g(x) = \uparrow$ then $f(g(x)) = \uparrow$. We say a total function f is **injective** or "1 to 1" if $f(x) = f(y)$ implies $x = y$; it is **surjective** or **onto** if $Y = \{f(x) : x \in X\}$; it is **bijective** if it is both injective and surjective.

The special functions of exponentiation $\exp_b(x)$ and logarithm $\log_b(x)$ to base $b > 0$ are more fully described in the Appendix of Chapter 2. Although these functions can be viewed as complex functions, we will exclusively treat them as real functions in this book. In particular, it means $\log_b(x)$ is undefined for $x \leqslant 0$. When the base b is not explicitly specified, it is assumed to be some constant $b > 1$. Two special bases[20] deserve their own notations: $\lg x$ and $\ln x$ refer to logarithms to base $b = 2$ and base $b = e = 2.718...$, respectively. In computer science, $\lg x$ is immensely useful. For any real a, we write $\log^a x$ as shorthand for $(\log x)^a$. E.g., $\log^2 x = (\log x)^2$. For any natural number i, let $\log^{(i)} x$ denote the i-fold application of the log-function. E.g., $\log^{(2)} x = \log(\log x)) = \log\log x$ and $\log^{(0)} x = x$. In fact, this notation can be extended to any integer i, where $i < 0$ indicates the $|i|$-fold application of exp.

Natural extensions of functions: from the function $f : X \to Y$, we can immediately define the function $F : 2^X \to 2^Y$ where $F(A) = \{f(a) : a \in A\}$ for $A \in 2^X$. Similarly, we can define $\overline{f} : X^* \to Y^*$ where $\overline{f}(x_1 x_2 \cdots x_n) = f(x_1)f(x_2)\cdots f(x_n)$. We call F and $\overline{f}$ the **natural extensions** of f to sets and sequences, respectively. To avoid introducing new symbols like F and $\overline{f}$, and may continue to use "f" for natural extensions. E.g., if $f(n) = n^2$, we may write $f(0, 1, 2, 3) = (0, 1, 4, 9)$, and if $A = \{1, 2, 3\}$, then $f(A) = \{1, 4, 9\}$.

¶A.6 Logic.

We assume the student is familiar with Boolean (or propositional) logic. In Boolean logic, each variable A, B stands for a proposition that is either true or false. Boolean logic deals with Boolean combinations of such variables: $\neg A, A \vee B, A \wedge B$. We also let $A \Rightarrow B$ denote **logical implication**, defined as $\neg A \vee B$. In particular $A \Rightarrow B$ iff $\neg B \Rightarrow \neg A$.

But mathematical facts go beyond propositional logic. Here is an example[21] of a mathematical assertion $P(x, y)$ about the real variables x, y:

$$P(x, y): \text{ There exists a real } z \text{ such that either } x \geqslant y \text{ or } x < z < y. \tag{36}$$

[19] The notion of "unlabeled graph" is only meaningful in a physical drawing, but impossible to represent in a computer, which needs distinct labels. We can do this by using "canonical labels" such as $X_n = \{1, \ldots, n\}$. But we want a single Hasse diagram $H(X_n)$ to represent each $[H(X_n)]$. One convention to help establish this is to insist that if x covers y then $x > y$ (as integers). But additional conventions will be needed to ensure that each Hasse shape $[H]$ has a unique Hasse diagram on X_n. See Exercise.

[20] Of course $\ln x$ has the (well-deserved) appellation "natural logarithm", but $\lg x$ has no special name. I suggest calling it the "Computer Science logarithm".

[21] When we formalize the logical language of discussion, what is called "assertion" here is often called "formula".

Assertions contain variables: for example, $P(x, y)$ in (36) contains x, y, z. Each variable has a range, often implicit: the variables x, y, z range over real numbers. Each variable is either **quantified** (either by "for all" or "there exists") or **unquantified**. Alternatively, quantified (unquantified) variables are said to be **bound** (**free**). In our example $P(x, y)$, z is bound while x, y are free. It is conventional to display the free variables as functional parameters of an assertion. The symbol $\forall$ stands for "for all" and is called the **universal quantifier**. Likewise, the symbol $\exists$ stands for "there exists" and is called the **existential quantifier**. Assertions with no free variables are called **statements**. We can always convert an assertion into a statement by adding some prefix to quantify each of the free variables. Thus, $P(x, y)$ can be converted into statements such as in (38) or as in $(\exists x \in \mathbb{R})(\forall y \in \mathbb{R})[P(x, y)]$. In general, if A and B are statements, so are any Boolean combinations of A and B, such as $A \wedge B$ and $\neg A$ or $A \vee B$. However, all statements can be transformed into the form

In computer programs, bound variables are just local variables and free variables are global variables!

$$(Q_1 x_1)(Q_2 x_2) \cdots (Q_n x_n)\,[P(x_1, x_2, \ldots, x_n)] \tag{37}$$

where Q_i is the ith quantifier part, each x_i is a distinct variable and $P(x_1, \ldots, x_n)$ a predicate. Such a form, where all the quantifiers appear before the predicate part, is said to be in **prenex form**.

The assertion $P(x, y)$ in (36) happens to be **valid**, meaning it is true for all instantiations of the free variables x, y. That is, if we bound all the free variables in $P(x, y)$ by universal quantifiers, then we get a true statement.

$$(\forall x, y \in \mathbb{R})[P(x, y)]. \tag{38}$$

All mathematical assertions are of this nature, i.e., valid. It is said that mathematical truths are universal: truthhood does not allow exceptions. If an assertion $P(x, y)$ has exceptions, let $E(x, y)$ be an assertion that holds at all exceptional instances of x, y (but $E(x, y)$ might capture more than just the exceptional instances). Now, the new statement $P(x, y) \vee E(x, y)$ is valid. A trivial choice for $E(x, y)$ is $\neg P(x, y)$.

In the above discussion, we make the conventional assumption that when the variables in an assertion are instantiated, then the assertion is either true or false. But in discussing partial functions, we need to generalize this to the setting where for some instances of x, y, the assertion $P(x, y)$ might be undefined (neither true nor false). We call P a **partial assertion** (or partial predicate). The quantified form $(\forall x)P(x)$ is then true if for all x in the domain, either $P(x)$ is undefined or $P(x)$ is true; similarly, $(\exists x)P(x)$ is true if there is some x in the domain such that $P(x)$ is defined and true. This extends naturally to predicates with more than one free variable.

More generally, we can define **predicates** to refer to any function with a finite domain. The usual 2-valued predicates may be called **logical predicates**. In geometric computation, we prefer a three-valued predicate since most geometric and topological distinctions are based on signs of real values.

¶A.7 Proofs and Induction.

Constructing proofs or providing counter examples to mathematical statements is a basic skill to cultivate. Three kinds of proofs are widely used: (i) case analysis, (ii) induction, and (iii) contradiction.

A proof by case analysis is often a matter of patience. But sometimes a straightforward enumeration of the possibilities will yield too many cases; clever insights may be needed to compress the argument. Induction is sometimes mechanical as well but very complicated inductions may also arise (Chapter 2 treats induction). Proof by contradiction may have a creative element: formulating an assertion to be contradicted!

HINT: It is a good idea to indicate the assertion that is going to be contradicted. For instance, to prove that an integer function $f(n)$ satisfies $f(n) < n$ using a proof by contradiction, you may say "By way of contradiction, assume $f(n) \geqslant n$". If this is a mouthful, we suggest the short hand "BWOC, let $f(n) \geqslant n$". Use BWOC like the other beloved acronym WLOG in proofs.

LOL, IMHO: "Do we need another acronym?" (DWNAA)

In proofs by contradiction, you will need to routinely negate a logical statement. Let us first consider the simple case of propositional logic. Here, you basically apply what is called De Morgan's Law: if A are B are truth values, then $\neg(A \vee B) = (\neg A) \wedge (\neg B)$ and $\neg(A \wedge B) = (\neg A) \vee (\neg B)$. For instance suppose you want to contradict the proposition $A \Rightarrow B$. You need to first know that $A \Rightarrow B$ is the same as $(\neg A) \vee B$. Negating this by de Morgan's law gives us $A \wedge (\neg B)$.

Next consider the case of quantified logic. **De Morgan's law for quantifiers** is the following pair of logical equivalences

$$\begin{aligned} \neg(\forall x)P(x) &\equiv (\exists x)[\neg P(x)], \\ \neg(\exists x)P(x) &\equiv (\forall x)[\neg P(x)]. \end{aligned}$$

In other words, we can push a negation past a quantifier by flipping the type of the quantifier. Of course, if P itself is quantified in prenex form, we can further transform $\neg P(x)$ by repeated application of the rule: e.g.,

$$\neg(\forall x(\exists y)(\forall z)[P(x,y,z)] \equiv (\exists x)(\forall y(\exists z)[\neg P(x,y,z)]$$

A useful extension of De Morgan's law for quantifiers arise as follows: in most applications, we do not write a "bald" quantification (Qx) where Q is $\exists$ or $\forall$. Instant we "bound" the domain of x within the quantifier by writing "$(Qx \in D)$" for some set D. If x is a real number, we may bound x by expressions like $(Qx > 3.14)$ or $(Qx \leqslant x_0)$. Call $(Qx \in D)$ a **bounded quantifier**. It's interpretation is given by this translation:

$$(Qx \in D)[P(x)] \equiv (Qx)[x \in D \Rightarrow P(x)]$$

We then have the bounded form of De Morgan's law for quantifiers:

$$\begin{aligned} \neg(\forall x \in D)P(x) &\equiv (\exists x \in D)[\neg P(x)], \\ \neg(\exists x \in D)P(x) &\equiv (\forall x \in D)[\neg P(x)]. \end{aligned}$$

These laws remain valid even when P is a partial predicate. A useful place to exercise these rules is to do some proofs involving the asymptotic notation (big-Oh, big-Omega, etc). See Exercises.

A proof Π can be organized in a variety of ways, but perhaps the simplest format is a sequence of assertions, $\Pi = (A_1, A_2, \ldots, A_n)$ where each A_i is either known to be true or can be deduced from $A_1, \ldots, A_{i-1}$ using sound rules of deduction. Note that any tree-like deduction can be linearized into the form Π. We can indicate this progression as

$$((\cdots((\texttt{true} \Rightarrow A_1) \Rightarrow A_2) \Rightarrow \cdots \Rightarrow A_{n-1}) \Rightarrow A_n) \tag{39}$$

where '$\Rightarrow$' should be read as 'implies'. We can simply write (39) as

$$A_1 \Rightarrow A_2 \Rightarrow \cdots \Rightarrow A_n$$

with that understanding that it is $\Rightarrow$ is left-associative.[22] We usually regard A_n as the conclusion of the proof. This is the normal direction of proof, where we proceed from known to new or unknown assertions. Of course, this is probably not the order in which you discover the proof. The more natural direction is the reverse direction, writing

$$(A_n \Leftarrow (A_{n-1} \Leftarrow \cdots \Leftarrow (A_n \Leftarrow (A_1 \Leftarrow \texttt{true}))\cdots)) \tag{40}$$

where '$\Leftarrow$' should be read as 'provided'. Again, (40) may be written

$$A_n \Leftarrow A_{n-1} \Leftarrow \cdots \Leftarrow A_1$$

if we assume that $\Leftarrow$ is right-associative. The advantage of this reverse mode is that you begin from what is known or required, and reduce it to more elementary assertions and eventually to a trivial or known assertion A_1. We illustrate this kind of proofs in §II.14 (orders of growth).

¶A.8 Formal Languages.
An **alphabet** is a finite set Σ of symbols. A finite sequence $w = x_1x_2\cdots x_n$ of symbols from Σ is called a **word** or **string** over Σ; the **length** of this string is n and denoted[23] $|w|$. When $n = 0$, this is called the **empty string** or **word** and denoted with the special symbol ϵ. The set of all strings over Σ is denoted Σ^*. A **language** over Σ is a subset of Σ^*.

¶A.9 Graphs.
A **hypergraph** is a pair $G = (V, E)$ where V is any set and $E \subseteq 2^V$. We call elements of V **vertices** and elements of E **hyper-edges**. In case $E \subseteq \binom{V}{k}$, we call G a k-graph. The case $k = 2$ is important and is called a **bigraph** (or more commonly, **undirected graph**). A **digraph** or **directed graph** is $G = (V, E)$ where $E \subseteq V^2 = V \times V$. For any digraph $G = (V, E)$, its **reverse** is the digraph $G^{rev} = (V, E')$ where $(u, v) \in E$ iff $(v, u) \in E'$. In this book, the word "graph" shall refer to a bigraph or digraph; the context should make the intent clear. The edges of graphs are often written '(u, v)' or 'uv' where u, v are vertices. We will prefer[24] to denote edge-hood by the notation $u{-}v$. Of course, in the case of bigraphs, $u{-}v = v{-}u$.

Often a graph $G = (V, E)$ comes with auxiliary data, say d_1, d_2, etc. In this case we denote the graph by

$$G = (V, E; d_1, d_2, \ldots)$$

using the semi-colon to mark the presence of auxiliary data. For example:
(i) Often one or two vertices in V are distinguished. If $s, t \in V$ are distinguished, we might write $G = (V, E; s, t)$. This notation might be used in shortest path problems where s is the source and t is the target for the class of paths under consideration.
(ii) A "weight" function $W : E \to \mathbb{R}$, and we denote the corresponding weighted graph by $G = (V, E; W)$.
(iii) Another kind of auxiliary data is **vertex coloring** of G, i.e., a function $C : V \to S$ where S is any set. Then $C(v)$ is called the **color** of $v \in V$. If $|S| = k$, we call C a k-coloring. The **chromatic graph** is therefore given by the triple $G = (V, E; C)$. An **edge coloring** is similarly defined, $C : E \to S$.

We introduce terminology for some special graphs: A graph $G = (V, E)$ is called the **empty graph** if V is the empty set. But if E is empty, then $G = (V, E)$ is called the **trivial graph**. Hence empty graphs are necessarily trivial but not vice-versa. It is usually implicit that we consider only non-trivial graphs. $K_n = (V, \binom{V}{2})$ denotes the **complete graph** on $n = |V|$

[22] In an algebra with a binary operation $*$, if the expression $a * b * c * d$ is interpreted as $((a * b) * c) * d$, we say $*$ is left-associative. Likewise, $a * (b * (c * d))$ is the right-associative interpretation.
[23] This notation should not be confused with the absolute value of a number or the size of a set. The context will make this clear.
[24] When we write $u{-}v$, it is really an assertion that the (u, v) is an edge. So it is redundant to say "$u{-}v$ is an edge".

vertices. A **bipartite graph** $G = (V, E)$ is a digraph such that $E \subseteq V_1 \times V_2$ where $V = V_1 \uplus V_2$. It is common to write $G = (V_1, V_2, E)$ in this case. The special case $K_{m,n} := (V_1, V_2, V_1 \times V_2)$ is called the **complete bipartite graph** where $m = |V_1|$ and $n = |V_2|$. We also say that a bigraph G is **bipartite** when G does not contain an odd cycle.

Two graphs $G = (V, E), G' = (V', E')$ are **isomorphic** if there is some bijection $\phi : V \to V'$ such that $\phi(E) = E'$ (the notation $\phi(E)$ has the obvious meaning). Since isomorphism is clearly an equivalence relation, let $[G]$ denote the isomorphism class of a graph G. We shall call $[G]$ the **shape** of G.

If $G = (V, E), G' = (V', E')$ where $V' \subseteq V$ and $E' \subseteq E$ then we call G' a **subgraph** of G. In case E' is the restriction of E to the edges in V', *i.e.*, $E' = E \cap V' \times V'$, then we say G' is the subgraph of G **induced by** V', or G' is the **restriction** of G to V'. We may write $G|V'$ for G'.

A **path** (from v_1 to v_k) is a sequence $(v_1, v_2, \ldots, v_k)$ of vertices such that (v_i, v_{i+1}) is an edge. Thus, we may also denote this path as $(v_1{-}v_2{-}\cdots{-}v_k)$. Call v_1 and v_k the **start** and **terminus** of the path, or collectively, they are the **endpoints** of the graph. A path is **closed** if $v_1 = v_k$ and $k > 1$. Two closed paths are **cyclic equivalent** if the sequence of edges they pass through are the same up to cyclic reordering. A cyclic equivalence class of closed paths is called a **cycle**. The **length** of a path is one less than the number of vertices in sequence defining the path; it is equal to the number of edges in the path. The length of a cycle is just the length of any of its representative closed paths. A graph is **acyclic** if it has no cycles. Sometimes acyclic bigraphs are called **forests**, and acyclic digraphs are called **dags** ("directed acyclic graph"). *For bigraphs, we further require that that closed paths have the form* $(v_1{-}v_2{-}v_3{-}\cdots{-}v_k)$ *where* $v_1 = v_k$ *and* $v_{i-1} \neq v_{i+1}$ *for all* $i = 1, \ldots, k$. Without this requirement, each edge $u{-}v$ in a bigraph would rise to a closed path of the form (u, v, u); such a bigraph would be considered "cyclic" for the wrong reason.

Two vertices u, v are **connected** if there is a path from u to v, and a path from v to u. (Note that in the case of bigraphs, there is a path from u to v iff there is a path from v to u.) We shall say v **is adjacent to** u if $u{-}v$. Connectivity is a symmetric binary relation for all graphs; adjacency is also a symmetric binary relation for bigraphs. It is easily seen that connectivity is also reflexive and transitive. This relation partitions the set of vertices into **connected components**.

In a digraph, **out-degree** and **in-degree** of a vertex is the number of edges issuing (respectively) from and into that vertex. The **out-degree** (resp., **in-degree**) of a digraph is the maximum of the out-degrees (resp., in-degrees) of its vertices. The vertices of out-degree 0 are called **sinks** and the vertices of in-degree 0 are called **sources**. The **degree** of a vertex in a bigraph is the number of adjacent vertices; the **degree** of a bigraph is the maximum of degrees of its vertices.

See Chapter IV for more graph concepts.

¶A.10 Trees.

A connected acyclic bigraph is called a **free tree**. A digraph such that there is a unique source vertex (called the **root**) and all the other vertices have in-degree 1, is called[25] a **tree**. The sinks in a tree are called **leaves** or **external nodes** and non-leaves are called **internal nodes**. In general, we prefer a terminology in which the vertices of trees are called **nodes**. Thus there is a

[25] One can also define trees in which the sense of the edges are reversed: the root is a sink and all the leaves are sources. We will often go back and forth between these two view points without much warning. E.g., we might speak of the "path from a node to the root". While it is clear what is meant here, but to be technically correct, we ought to speak awkwardly of the path in the "reverse of the tree".

unique path from the root to each node in a tree. If u, v are nodes in T then u is a **descendant** of v if there is a path from v to u. Every node v is a descendant of itself, called the **improper descendant** of v. All other descendants of v are called **proper**. We may speak of the **child** or **grandchild** of any node in the obvious manner. The **degree** of v is the number of children of v. The **degree** of a tree is the maximum degree of any of its nodes. The reverse of the descendant binary relation is the **ancestor** relation; thus we have **proper ancestors**, **parent** and **grandparent** of a node.

The **subtree** at any node u of T is the subgraph of T obtained by restricting to the descendants of u. The **size** of T is the number of nodes in T. The **depth** of a node u in a tree T is the length of the path from the root to u. So the root is the unique node of depth 0. The **depth of** T is the maximum depth of a node in T. The **height** of a node u is just the depth of the subtree at u; alternatively, it is the length of the longest path from u to its descendants. Thus u has height 0 iff u is a leaf iff u has no children. The collection of all nodes at depth i is also called the i**th level** of the tree. Thus level zero is comprised of just the root. The **height** of T is just the height of its root. NOTE: these concepts are clear if T has size $\geqslant 1$. If the size of T is 0, we define its height and depth to be -1.

We normally draw a tree with the root at the top of the figure, and edges are implicitly direction from top to bottom. Some additional tree concepts may be found in Chapter III, especially those related to binary trees.

¶A.11 Programs and Pseudo-Code.

In this book, we present algorithms in an informal under-specified programming language that combines mathematical notations with standard programming language constructs, sprinkled with English. Conventionally, we call such program descriptions **pseudo-code** *The goal is pseudo-code is human understanding of the algorithmic logic.* This includes expose the underlying algorithmic logic and basic flow of control. The goal does not including producing a compilable code in any particular conventional programming language! Nevertheless, it is often easy to transcribe pseudo-code into compilable code in real programming languages such as `C++` or `Java`. There are two good reasons for preferring pseudo-code:

- It exploits natural language and is thus easier to understand. Simultaneously, it exploits mathematical language which is universal and very compact.
- Being informal, pseudo-code can flexibly be written to provide as much (or as little) detail as is necessary for a target audience, or for a specific pedagogical purpose.

> Programming languages are harder to understand because they are intended for machine consumption, and that could get in the way of human understanding. The advantage of writing compilable code is that it could be given to a computer for execution. Unfortunately, the "half-life" of programming languages tend to be rather short compared to that of natural languages or mathematical language. Informally, say the half-life of a programming language is the average time it takes before some program in your library collection will no longer compile. My (uneducated) guess is a half-life of $1-5$ years, depending on the language (e.g., the `C` language would have a much longer half-life than `C++`). Similarly, the half-life of pseudo code is the average time before most people in the field find some program in your pseudo-code collection impossible to understand. Another uneducated guess is a half-life of a human generation (30 years).

For the purposes of this book, pretend that we have a "pseudo Programming Language" called **pseudo-PL** that has a few additional guidelines:

pseudo-PL is appropriately amorphous by design

- We use standard programming constructs such as if-then-else, while-loop, return statements, etc.

no clutter language

- To reduce clutter, we indicate the structure of programming blocks by indentation and newlines only. In particular, we avoid explicit block markers such as "begin...end" or"{...}".
- Single line comments in a program are indicated in two ways:
  ▷ *This is a forward comment*
  ◁ *This is a backward comment*
  These comments either precede (if a forward comment) or follows (if a backward comment) the code that it describes. We have little need for multiline comments in pseudo-PL because the code is supplemented by off-line explanations to the same purpose.
- Programming variables are undeclared, and implicitly introduced through their first use. They are not explicitly typed, but the context should make this clear. This is in the spirit of modern scripting languages such as `Perl`, and consistent with our clutter-free pseudo-code.
- Normally, each line is a command, so we may terminate a command line with the traditional semicolon (;) or without any punctuation marks if it is clear. If the line is more "Englishy", we prefer to terminate in a full stop. But if we put two or commands on one line, we generally separate them with semicolons. What if a command needs more than one line? In many computer languages, the continuation symbol is \. But in order to produce more human friendly programs, we could use ellipsis "..." at the end of a line to indicate its continuation to the next line. But if the line is an English sentence, we can even drop the ellipsis and indent the continuation line appropriately.

Programmers use '=' for assignment and '==' for equality test. We opt to preserve the mathematical meaning of '='.

- Informally, the equality symbol '=' is often overloaded to indicate the assignment operator as well as the equality test. We will use ← for assignment operator, and reserve '=' for equality test.
- In the style of `C` or `Java`, we write "x++" (resp., "++x") to indicate the increment of an integer variable x. The value of this expression is the value of x before (resp., after) incrementing. There is an analogous notation for decrementing, x-- and --x.

Here is a recursive program written in pseudo-PL to compute the Factorial function:

```
FAC(n):
 Input: natural number n.
 Output: n!
 ▷ Base Case
1.    If n ⩽ 1 Return(1)
 ▷ General Case
2.    Return(n · FAC(n − 1))     ◁ This is a recursive call
```

¶A.12 How to answer algorithmic exercises.

In our exercises, whenever we ask you to give an algorithm, it is best to write in pseudo-code. We suggest that you emulate our pseudo-PL form of presentation. Students invariably ask about the level of detail in pseudo-code. The general answer is *as much detail as one needs to know how to reduce it to compilable programs in conventional programming languages.* Here is a checklist you can use:

- *Specify your input and output.* This cannot be emphasized enough. We cannot judge your algorithm if we do not know what to expect from its output! *sine qua non!*
- *Take advantage of well-known algorithms.* For instance, if you need to to sort, you should generally be able to just[26] invoke a suitable sorting routine.
- *Reduce all operations to $\mathcal{O}(1)$ time operations.* Do this when Rule 1 does not apply. Sometimes, achieving $\mathcal{O}(1)$ time may depend on a suitable choice of data structures. If so, be sure to explain this.
- *Use progressive algorithm development.* Even pseudo-code may be incomprehensible without a suitable orientation – it is never wrong to precede your pseudo-code with some English explanation of what the basic idea is. In more complicated situations, do this in 3 steps: explain basic ideas, give pseudo-code, further explain certain details in the pseudo-code.
- *Use standard algorithmic paradigms.* In this book, we will introduce well-known paradigms such as divide-and-conquer, greedy methods, dynamic programming, etc. Another important paradigm is the notion of "shell-programming" i.e., driver programs; See tree and graph traversals in Chapters III and IV. *NOT "shell" in the sense of bash, csh, tcsh, etc.*
- *Explain and initialize all variables and data structures.* Most non-trivial algorithms have some data structures, possibly the humble array. Critical variables (counters, coloring schemes) ought to be explained too. You must show how to initialize them.
- *The control structure of your algorithms should be evident.* All the algorithms you design should have simple control structures – typically a simple loop or a doubly-nested loops. Triply-nested loops do arise (e.g., dynamic programming) but deeper nesting is seldom needed. Each loop should use standard programming constructs (for-loop, while-loop, do-loop, etc). It is an axiom[27] that if a problem can be solved, then it is solvable by clean loop structures.
- *Correctness.* This is an implicit requirement of all algorithms. All the algorithms we study must halts on all (valid) inputs. Correctness of such algorithms is traditionally split into two distinct requirements:
  (1) The algorithm halts.
  (2) The output is correct when it halts. This is also known as **partial correctness**.
  A simple method to prove partial correctness is this: at the beginning of each iteration of a loop, you should be able to attach a suitable **invariant** (called **assertion** in standard programming languages). Partial correctness follows easily if the appropriate invariants hold.

  meta principle: *every loop has an invariant – you just have to look for it*
- *Analysis and Efficiency.* This is a more advanced requirement. But since this is what algorithmics is about, we view it as part and parcel of any algorithm in this book. You should always be able to give a big-Oh analysis of your algorithm. In most cases, non-polynomial time solutions are unnecessarily inefficient.

[26] In computing, this is known as "code reuse". Others call it "not reinventing the wheel".
[27] There are theorems about the universality of loop-programs (Meyer and McCreight) and the possibility of avoiding "go-to" statements.

¶A.13 LIST OF ABBREVIATIONS

- AVL: Adel'son-Vel'skii and Landis tree (§III.6)
- BFS: Breadth First Search (§IV.4)
- BST: Binary search Tree (§III.3)
- DFS: Depth First Search (§IV.5)
- `Grd`: Greedy Algorithm for linear bin packing
- `FF`: First Fit Algorithm for bin packing
- iff: if and only if
- MST: Minimum Spanning Tree
- TSP: Traveling Salesman Problem
- WLOG or wlog: Without loss of generality

Exercises

Exercise 10.5: Please write a piece of Java code that has complexity $O(m(n^2 + p))$ where m, n, p are positive integers. ♢

Exercise 10.6: Show that if x is real and a a positive integer, then $\lfloor x/a \rfloor = \lfloor \lfloor x \rfloor /a \rfloor$. Similarly, $\lceil x/a \rceil = \lceil \lceil x \rceil /a \rceil$. ♢

Exercise 10.7: The following is a useful result about iterated floors and ceilings.
(a) Let $x > 0$ be real and $b > 1$ be an integer. Let $N_0 := \lfloor x \rfloor$ and for $i \geqslant 0$, $N_{i+1} := \lfloor N_i/b \rfloor$. Show that $N_i = \lfloor x/b^i \rfloor$. Similarly for ceilings.
(b) Let $u_0 = 1$ and $u_{i+1} = \lfloor 5u_i/2 \rfloor$ for $i \geqslant 0$. Show that for $i \geqslant 4$, $0.76(5/2)^i < u_i \leqslant 0.768(5/2)^i$. HINT: $r_i := u_i(2/5)^i$ is non-increasing; give a lower bound on r_i ($i \geqslant 4$) based on r_4.
(c) Show that if b is not an integer, part(a) may fail. ♢

Exercise 10.8: Let $\mu > 1$ be integer and $0 < \rho < 1$ a real number. Think of μ as arbitrary but ρ as fixed. We want the integer expression $E(\mu)$ such that

$$E(\mu) > \mu\rho \geqslant E(\mu) - 1.$$

If ρ is irrational, then $E(\mu) = \lceil \mu\rho \rceil$ is the desired expression. What if $\rho = p/q$ is a rational? ♢

Exercise 10.9: It is well known that $n = \lfloor n/2 \rfloor + \lceil n/2 \rceil$ when n is integer. When n is non-integer, show that

$$n - 1 < \lfloor n/2 \rfloor + \lceil n/2 \rceil < n + 1$$

and this is the sharpest possible bound. ♢

Exercise 10.10: In a certain household, the laundry is sorted into two bins with the labels *Whites and Underclothing Only*, and *All Others*. What domestic catastrophe ensues (this did happen) when two family members interprets the signs in completely different ways? HINT: Both "White" are "Underclothing" can be interpreted as predicates or as nouns. ♢

Exercise 10.11: Consider the following sentence:

$$(\forall x \in \mathbb{Z})(\exists y \in \mathbb{R})(\exists z \in \mathbb{R}) \left[(x > 0) \Rightarrow ((y < x < y^{-1}) \wedge (z < x < z^2) \wedge (y < z)) \right] \tag{41}$$

Note that the range of variable x is $\mathbb{Z}$, not $\mathbb{R}$. This is called a **universal sentence** because the leading quantifier is the universal quantifier $(\forall)$. Similarly, we have **existential sentence**.

(i) Negate the sentence (41), and then apply De Morgan's law to rewrite the result as an existential sentence.

(ii) Give a counter example to (41).

(iii) By changing the clause "$(x > 0)$", make the sentence true. Indicate why it would be true.

♢

Exercise 10.12: Let $f(n) = (\log n)^{\log n}$. Suppose you want to prove

$$f(n) \neq \mathcal{O}(f(n/2)).$$

(a) Using de Morgan's law, show that this amounts to saying that for all $C > 0, n_0$ there exists n such that

$$(n \geqslant n_0) \wedge f(n) > Cf(n/2).$$

(b) Complete the proof by finding a suitable n for any given C, n_0. ♢

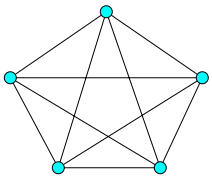
Exercise 10.13: What is the connection between the concept of a bipartite digraph and a bipartite bigraph? ♢

Exercise 10.14: Let G be a planar graph with $v \geqslant 3$ vertices and e edges. Prove the inequality: $e \leqslant 3v - 6$.
HINT: Let $em(G)$ be an embedding of G in the plane where $em(v) \in \mathbb{R}^2$ and $em(u–v)$ is a curve connecting $em(u)$ and $em(v)$, with no two curves intersecting each other (except possibly at their endpoints). These curves enclose regions of the plane called **faces**: let there be f faces. Among these faces, there is a unique unbounded face F_∞. For a fixed n, the number of edges e is maximized when each face is bounded by 3 curves (even F_∞). What is the relationship between v, e, f in this case? ♢

Exercise 10.15: The previous exercise shows that *a planar graph on n vertices has at most $3n-6$ edges.* Let us restate it as follows:

$$(G \text{ is a planar graph and has } n \text{ vertices}) \Rightarrow (G \text{ has } \leqslant 3n-6 \text{ edges}).$$

(i) State the contra-positive of this statement.
(ii) The complete graph on 5 vertices is denoted by K_5. Prove that K_5 is not planar by using the contra-positive statement. ♢



Complete graph K_5

Exercise 10.16: The road map of a certain state has 314, 159 road segments, where each road segment connects a pair of distinct junctions. Assume that every road segment connect a different pair of junctions. What is the least number of junctions in that state? HINT: use the theorem in the previous exercise about planar graph. ♢

Exercise 10.17: Let $P : D \to \{0, 1\}$ be a partial predicate over some domain D. When we do quantification, $\forall x$ and $\exists x$ it is assumed that x range over D. Show that the following equivalences (called "de Morgan's laws for quantifiers") hold:
(a) $\neg(\forall x)P(x)$ is equivalent to $(\exists x)\neg P(x)$
(b) $\neg(\exists x)P(x)$ is equivalent to $(\forall x)\neg P(x)$ ♢

Exercise 10.18: We say that a tree is **ordered** if the children of each node are totally ordered (so we can speak of the first child, the last child, etc). What is wrong with defining a **binary tree** as a rooted, ordered tree of degree 2? ♢

Exercise 10.19: Prove these basic facts about binary trees: assume $n \geqslant 1$.
(a) A full binary tree on n leaves has $n-1$ internal nodes.
(b) Show that every binary tree on n nodes has height at least $\lceil \lg(1+n) \rceil - 1$. HINT: define $M(h)$ to be the maximum number of nodes in a binary tree of height h.
(c) Show that the bound in (b) is tight for each n.
(d) Show that a binary tree on $n \geqslant 1$ leaves has height at least $\lceil \lg n \rceil$. HINT: use a modified version of $M(h)$.
(e) Show that the bound in (d) is tight for each n. ♢

Exercise 10.20: (Erdös-Rado) Show that in any 2-coloring of the edges of the complete graph K_n, there is a monochromatic spanning tree of K_n. HINT: use induction. ♢

Exercise 10.21: Let T be a binary tree on n nodes.
(a) What is the minimum possible number of leaves in T?
(b) Show by strong induction on the structure of T that T has at most $\lfloor \frac{n+1}{2} \rfloor$ leaves. This is an exercise in case analysis, so proceed as follows: first let n be odd (say, $n = 2N+1$) and assume T has $k = 2K+1$ children in the left subtree. There are 3 other cases.
(c) Give an alternative proof of part (b): show the result for n by a weaker induction on $n-1$ and $n-2$.
(d) Show that the bound in part (b) is the best possible by describing a T with $\lfloor \frac{n+1}{2} \rfloor$ leaves. HINT: first show it when $n = 2^t - 1$. Alternatively, consider binary heaps. ♢

Exercise 10.22:
(a) A binary tree with a key associated to each node is a binary search tree iff the in-order

listing of these keys is in increasing order.
(b) Given *both* the post-order and in-order listing of the nodes of a binary tree, we can reconstruct the tree. ♢

Exercise 10.23: Prove that if R is a partial order of X, then the Hasse diagram $H(R)$ of R has at most $|X| - 1$ edges. ♢

End Exercises

§11 APPENDIX B: The RAM Model

Comparison tree models are somewhat special because they are non-uniform. We now present a uniform computational model called the **Random Access Memory model** (RAM model, for short).

¶B.0 Universal Computational Models.
A well-known universal model is the Turing machine. A model is **universal** if every solvable problem can be solved by programs in the universal model. Universality seems like a tall order; indeed, universality is not established by a theorem but by agreement among experts! This agreement is called **Church's Thesis**. But once we agree to call one model "universal", then the universality of others can[28] be established by theorems. To achieve the power of universality, the data structures in the universal model must be general enough to encode any other data structure or data objects.

¶B.1 Register Model.
The RAM model can be viewed as an abstract form of a simple assembly language. In turn, the RAM model is a generalization of an even simpler model called the **Register model**. So we shall begin by describing such "register machines".

(a) Data objects. We assume infinitely many **(storage) registers**, each indexed by an integer $0, \pm 1, \pm 2, \pm 3, \ldots$. Register 0 is special and is known as the **accumulator**. Each register can store an integer. The integer can be arbitrarily large.

this is not a 64-bit or even 128-bit machine...

(b) Primitive Operations. In the simplest form, each operation has 2 fields:

$$\langle OPERATOR \rangle \langle ARG \rangle$$

There is one argument $\langle ARG \rangle$ whose nature is determined by the $\langle OPERATOR \rangle$: $\langle ARG \rangle$ is either an integer (denoted n below) or a label (denoted ℓ below). But in general, an operation can have up to 4 fields:

$$[\langle LABEL \rangle :] \langle OPERATOR \rangle \langle ARG \rangle [\langle COMMENTS \rangle]$$

where $\langle LABEL \rangle$ and $\langle COMMENTS \rangle$ are arbitrary non-empty alphabetic strings – the square brackets indicate that these are optional fields. The contents of register n is a number, denoted $c(n)$. Thus the **contents function** has the form $c : \mathbb{Z} \to \mathbb{Z}$. The operators, their arguments and actions are specified in Table 1.

Most of these operations have the obvious meaning. For the DIV operation, we assume that the integer quotient is put into $c(0)$ and any remainder is discarded.

[28] Technically, by "encoding and simulation".

(c) Semantics. A **register program** P is any finite sequence of such primitive operations. There is also a **label** function λ which, for any label ℓ, returns the index $\lambda(\ell)$ of the instruction in this sequence P. Now we can define a **computation** in which at each instance, we have a **program counter** whose value is the index of the (current) instruction in P being executed. When we execute the current instruction, this results in a transformation of the contents function. This transformation is specified by the last column of Table 1. For instance, "GET 4" will put update the value of $c(0)$ to be the value $c(4)$, but no other register contents are changed. Subsequently, transformations are defined by successive instructions of the program (this means that the program counter is simply incremented). The exception is when there is a successful jump to some label ℓ, in which case the program counter is updated to $\lambda(\ell)$. The computation halts on encountering the HALT operation, or when there is no "next" instruction, or upon jumping to some non-existent label.

Operator	Argument	Semantics
GET	n	$c(0) \leftarrow c(n)$.
PUT	n	$c(n) \leftarrow c(0)$.
ZERO	n	$c(n) \leftarrow 0$.
INC	n	$c(n) \leftarrow c(n) + 1$.
ADD	n	$c(0) \leftarrow c(0) + c(n)$.
SUB	n	$c(0) \leftarrow c(0) - c(n)$.
MUL	n	$c(0) \leftarrow c(0) \times c(n)$.
DIV	n	$c(0) \leftarrow c(0) \div c(n)$, error if $c(n) = 0$.
JUMP	ℓ	Go to label ℓ.
JPOS	ℓ	If $c(0) > 0$ then go to ℓ.
JNEG	ℓ	If $c(0) < 0$ then go to ℓ.
HALT		The computation halts.

Table 1: Instruction Set for Register Machines

(d) Input/Output conventions. We assume that a finite number of registers is initialized with an encoding of the input, while the rest of the registers are initially zero. The convention for the output is some simple function of the final contents function. E.g., the output may be defined to be $c(0)$. As an exercise, the reader may write a RAM program to compute the maximum of three numbers. Use the convention that the input numbers are in registers $1, 2, 3$ and the maximum value must be output in register 0).

¶B.2 Random Access Feature.

It is a very small step to turn the above Register Model into a RAM Model. First notice that a register program can only access a fixed set of registers. In order to allow a program to access an arbitrary number of registers, we introduce "indirect addressing", a concept from computer architecture. We allow another form of integer argument, denoted $@n$ where $n \in \mathbb{N}$. This means we are using the value $c(c(n))$ instead of $c(n)$ as the actual argument for the operation. Thus $c(n)$ is interpreted as the address of the actual argument. We call $@n$ an "indirect argument". For instance, "GET @4" results in the assignment $c(0) \leftarrow c(c(4))$. If register 4 contains 256, and register 256 contains -1, this means $c(0)$ is assigned the value -1. Similarly, "PUT @4" will place the value $c(0)$ into register 256. Thus, a **RAM program** is basically a register program in which we allow indirect addressing.

¶B.3 Extensions.

By design, the instruction set of our RAM is parsimonious and primitive. There are many possible extensions where we enrich the instruction set; these makes programming convenient, but do not extend the power of the model. For instance, we can allow another kind of integer argument denoted "$= n$". This means that the value n itself is being used – we never have to access the contents function c. This is called a **literal argument**. Literal arguments are useful for GET and the arithmetic instructions, but meaningless for PUT instruction.

The above model may also be called the **Integer RAM model**. This can be generalized

to the **Real RAM model** where where the registers can store an arbitrary real number, and possibly augmenting the primitive operations with other real functions (such as computing square roots). When we use a register n for indirect addressing ($@n$), we need to have some convention for handling the case where its contents $c(n)$ is not an integer. However, we must be aware that the Real RAM is a highly non-trivial extension; it is a useful model but it is not realistic by any stretch of imagination. One reason is that the set of real numbers are uncountable and we have no hope of representing all real numbers in a finitistic scheme.

We can also augment the model with new primitive operations. For instance, to write programs in the Tape Model (see ¶9), we just have to add operations corresponding to the `READ`, `WRITE`, `RESET` in (7), and the `EOT` test.

¶B.4 Higher Level Languages and Universality.

In practice, we may write our program using a more abstract language or a "higher level" language. Thus, we may use well-known constructs such as for-loops and if-then-else, and even allow recursion. Nevertheless, such extensions of the model can be translated into a standard RAM program. In this sense, the RAM model is universal in the sense that there is no computational model that is more powerful. In complexity theory, this claim about "universality" is called **the Church-Turing Thesis**.

EXERCISES

Exercise 11.24: Argue that the (Integer) RAM model is equivalent to the Turing model. HINT: This amounts to showing that for any Turing machine M (resp., any RAM program P), there is a RAM Program P' (resp., a Turing machine M') such that P' simulates M (resp., M' simulates P). ♢

Exercise 11.25: Recall the Merge Algorithm described in ¶14. Please convert it into a RAM program. In other words, you must use the instruction set in Table 1. ♢

Exercise 11.26: Write RAM algorithms for the following problems:

(a) Sort a sequence of input numbers.

(b) Compute the GCD of two integers.

(c) Solve the ranking problem in ¶4. You need two algorithms, for preprocessing and for answering queries.

(d) Solve the dynamic ranking problem in ¶5. You need four algorithms here: to initialize an empty data structure, to insert, to delete, and to answer rank queries.

Be sure to state your input/output conventions. Also, describe any data structures you use. ♢

Exercise 11.27: Reduce the number of primitive operators (listed in table above) for our RAM model to a minimum. In particular, show that we only need the following:

$$GET, PUT, ZERO, INC, JPOS, HALT$$

Show that a RAM model with this set of instructions can simulate our original RAM model. ♢

Exercise 11.28: Show how to implement higher level language constructs such as *for-loops, if-then-else, case-statements* in our RAM model. ♢

End Exercises

References

[1] G. E. Collins. The computing time analsysis of the Euclidean algorithm. *SIAM J. Computing*, 1:1–17, 1974.

[2] T. H. Corman, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press and McGraw-Hill Book Company, Cambridge, Massachusetts and New York, second edition, 2001.

[3] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin, revised 3rd edition edition, 2008.

[4] D. G. Kirkpatrick and R. Seidel. The ultimate planar convex hull algorithm? *SIAM J. Comput.*, 15:287–299, 1986.

[5] D. E. Knuth. *The Art of Computer Programming: Sorting and Searching*, volume 3. Addison-Wesley, Boston, 1972.

[6] D. E. Knuth. *The Art of Computer Programming: Fundamental Algorithms*, volume 1. Addison-Wesley, Boston, 2nd edition edition, 1975.