**SEARCHING, SORTING, AND ASYMPTOTIC COMPLEXITY**

**Lecture 10**

**Mergesort**

```java
/** Sort b[h..k]. Precondition: b[h..t] and b[t+1..k] are sorted. */
public static merge(int[] b, int h, int t, int k) {
    Copy b[h..t] into another array c;
    Copy values from c and b[t+1..k] in ascending order into b[h..]
}
```

Merge two adjacent sorted segments

We leave you to write this method. It is not difficult. Just have to move values from c and b[t+1..k] into b in the right order, from smallest to largest. Runs in time \(O(k+1-h)\)

**QuickSort versus MergeSort**

```java
/** Sort b[h..k] */
public static void QS(int[] b, int h, int k) {
    if (k - h < 1) return;
    int j = partition(b, h, k);
    QS(b, h, j-1);
    QS(b, j+1, k);
}
```

```java
/** Sort b[h..k] */
public static void MS(int[] b, int h, int k) {
    if (k - h < 1) return;
    MS(b, h, (h+k)/2);
    MS(b, (h+k)/2 + 1, k);
    merge(b, h, (h+k)/2, k);
}
```

One processes the array then recurses. One recurses then processes the array.

**Readings, Homework**

- **Textbook:** Chapter 4
- **Homework:**
  - Recall our discussion of linked lists and A2.
  - What is the worst case complexity for appending an items on a linked list? For testing to see if the list contains \(X\) ? What would be the best case complexity for these operations?
  - If we were going to talk about complexity (speed) for operating on a list, which makes more sense: worst-case, average-case, or best-case complexity? Why?
What Makes a Good Algorithm?

Suppose you have two possible algorithms or ADT implementations that do the same thing; which is better?

What do we mean by better?
- Faster?
- Less space?
- Easier to code?
- Easier to maintain?
- Required for homework?

How do we measure time and space of an algorithm?

Basic Step: One “constant time” operation

- If-statement: number of basic steps on branch that is executed
- Loop: (number of basic steps in loop body) * (number of iterations) – also bookkeeping
- Method: number of basic steps in method body (include steps needed to prepare stack-frame)

Counting basic steps in worst-case execution

** Linear Search **

```java
/** return true iff v is in b */
static boolean find(int[] b, int v) {
    for (int i = 0; i < b.length; i++) {
        if (b[i] == v) return true;
    }
    return false;
}
```

Let n = b.length

<table>
<thead>
<tr>
<th>worst-case execution</th>
<th>basic step</th>
<th># times executed</th>
</tr>
</thead>
<tbody>
<tr>
<td>i=0;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>i &lt; b.length</td>
<td>n+1</td>
<td></td>
</tr>
<tr>
<td>i++</td>
<td>n</td>
<td></td>
</tr>
<tr>
<td>b[i] == v</td>
<td>n</td>
<td></td>
</tr>
<tr>
<td>return true</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>return false</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3n + 3</td>
<td></td>
</tr>
</tbody>
</table>

We sometimes simplify counting by counting only important things. Here, it’s the number of array element comparisons b[i] == v, that’s the number of loop iterations: n.

** Sample Problem: Searching **

** Second solution: Binary Search **

```java
/** b is sorted. Return h satisfying b[0..h] <= v < b[h+1..] */
static int bsearch(int[] b, int v) {
    int h= -1;
    int k= b.length;
    while (h+1 != k) {
        int e= (h+ k)/2;
        if (b[e] <= v)  h= e;
        else
            k= e;
    }
    return h;
}
```

Number of iterations (always the same):

- log b.length
- Therefore, log b.length array comparisons

Definition of \( O(\)…\)

Formal definition: \( f(n) \text{ is } O(g(n)) \) if there exist constants \( c \) and \( N \) such that for all \( n \geq N \), \( f(n) \leq c \cdot g(n) \)

Graphical view

- Get out far enough for \( n \geq N \)
- \( c \cdot g(n) \) is bigger than \( f(n) \)
What do we want from a definition of “runtime complexity”?

Formal definition: \( f(n) \) is \( O(g(n)) \) if there exist constants \( c \) and \( N \) such that for all \( n \geq N \), \( f(n) \leq c \cdot g(n) \).

Roughly, \( f(n) \) is \( O(g(n)) \) means that \( f(n) \) grows like \( g(n) \) or slower, to within a constant factor.

Prove that \( (n^2 + n) \) is \( O(n^2) \)

Formal definition: \( f(n) \) is \( O(g(n)) \) if there exist constants \( c \) and \( N \) such that for all \( n \geq N \), \( f(n) \leq c \cdot g(n) \).

Example: Prove that \( (n^2 + n) \) is \( O(n^2) \)

\[
\begin{align*}
f(n) &= \text{\(<\text{definition of } f(n)\rangle\)} \\
&= n^2 + n \\
&\leq \text{\(<\text{for } n \geq 1\rangle\)} \\
&= n^2 + n \\
&= \text{\(<\text{arith}\rangle\)} \\
&= 2 \cdot n^2 \\
&= \text{\(<\text{choose } g(n) = n^2\rangle\)} \\
&= 2 \cdot g(n)
\end{align*}
\]

Choose \( N = 1 \) and \( c = 2 \).

O(…) Examples

Let \( f(n) = 3n^2 + 6n - 7 \)

- \( f(n) \) is \( O(n^2) \)
- \( f(n) \) is \( O(n^3) \)
- \( f(n) \) is \( O(n^4) \)
- ...

\[
\begin{align*}
p(n) &= 4 \cdot n \cdot \log n + 34 \cdot n - 89 \\
p(n) &= \text{\(O(n \log n)\)} \\
p(n) &= \text{\(O(n^2)\)} \\
h(n) &= 20 \cdot 2^n + 40n \\
h(n) &= \text{\(O(2^n)\)} \\
a(n) &= 34 \\
a(n) &= \text{\(O(1)\)}
\end{align*}
\]

Only the leading term (the term that grows most rapidly) matters.

If it’s \( O(n^2) \), it’s also \( O(n^3) \) etc! However, we always use the smallest one.

Problem-size examples

Suppose a computer can execute 1000 operations per second; how large a problem can we solve?

<table>
<thead>
<tr>
<th>alg</th>
<th>1 second</th>
<th>1 minute</th>
<th>1 hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O(n) )</td>
<td>1000</td>
<td>60,000</td>
<td>3,600,000</td>
</tr>
<tr>
<td>( O(n \log n) )</td>
<td>140</td>
<td>4893</td>
<td>200,000</td>
</tr>
<tr>
<td>( O(n^2) )</td>
<td>31</td>
<td>244</td>
<td>1897</td>
</tr>
<tr>
<td>( 3n^2 )</td>
<td>18</td>
<td>144</td>
<td>1096</td>
</tr>
<tr>
<td>( O(n^3) )</td>
<td>10</td>
<td>39</td>
<td>153</td>
</tr>
<tr>
<td>( O(2^n) )</td>
<td>9</td>
<td>15</td>
<td>21</td>
</tr>
</tbody>
</table>

Commonly Seen Time Bounds

\[
\begin{array}{|c|c|c|}
\hline
O(1) & \text{constant} & \text{excellent} \\
O(\log n) & \text{logarithmic} & \text{excellent} \\
O(n) & \text{linear} & \text{good} \\
O(n \log n) & n \log n & \text{pretty good} \\
O(n^2) & \text{quadratic} & \text{OK} \\
O(n^3) & \text{cubic} & \text{maybe OK} \\
O(2^n) & \text{exponential} & \text{too slow} \\
\hline
\end{array}
\]
Worst-Case/Expected-Case Bounds

- May be difficult to determine time bounds for all imaginable inputs of size n

Simplifying assumption #4: Determine number of steps for either
- worst-case or
- expected-case or
- average case

- Worst-case
  - Determine how much time is needed for the worst possible input of size n
- Expected-case
  - Determine how much time is needed on average for all inputs of size n

Simplifying Assumptions

- Use the size of the input rather than the input itself – n
- Count the number of “basic steps” rather than computing exact time
- Ignore multiplicative constants and small inputs (order-of, big-O)
- Determine number of steps for either
  - worst-case
  - expected-case

These assumptions allow us to analyze algorithms effectively

Worst-Case Analysis of Searching

**Linear Search**
// return true iff v is in b
static boolean find(int[] b, int v) {
for (int x : b) {
  if (x == v) return true;
}
return false;
}
worst-case time: O(n)

**Binary Search**
// Return h that satisfies
// b[0..h] <= v < b[h+1..]
static boolean bsearch(int[] b, int v {
  int h= -1; int t= b.length;
  while ( h != t-1 ) {
    int e= (h+t)/2;
    if (b[e] <= v)  h= e;
    else t= e;
  }
}
Always takes ~\(\log n+1\) iterations.
Worst-case and expected times: \(O(\log n)\)

Comparison of linear and binary search

Analysis of Matrix Multiplication

Multiply n-by-n matrices A and B:

- Input size is really \(2n^2\), not \(n\)
- Worst-case time: \(O(n^3)\)
- Expected-case time: \(O(n^2)\)

For (i = 0; i < n; i++)
  for (j = 0; j < n; j++) {
    c[i][j] = 0;
    for (k = 0; k < n; k++)
      c[i][j] += a[i][k]*b[k][j];
  }
Remarks

Once you get the hang of this, you can quickly zero in on what is relevant for determining asymptotic complexity
- Example: you can usually ignore everything that is not in the innermost loop. Why?

One difficulty:
- Determining runtime for recursive programs
  
  Depends on the depth of recursion

Why bother with runtime analysis?

Computers so fast that we can do whatever we want using simple algorithms and data structures, right?
Not really – data-structure/algorithm improvements can be a very big win

Scenario:
- A runs in $n^2$ msec
- A' runs in $n^2/10$ msec
- B runs in $10n \log n$ msec

Problem of size $n = 10^3$
- A: $10^6$ sec $\approx$ 17 minutes
- A': $10^5$ sec $\approx$ 1.7 minutes
- B: $10^6$ sec $\approx$ 1.7 minutes

Problem of size $n = 10^6$
- A: $10^9$ sec $\approx$ 30 years
- A': $10^8$ sec $\approx$ 3 years
- B: $2 \times 10^5$ sec $\approx$ 2 days

1 day $= 86,400$ sec $\approx 10^5$ sec
1,000 days $\approx$ 3 years

Algorithms for the Human Genome

Human genome
- 3.5 billion nucleotides
- $\sim$ 1 Gb

@1 base-pair instruction/1 sec
- $n^2$ $\rightarrow$ 388445 years
- $n \log n$ $\rightarrow$ 30.824 hours
- $n$ $\rightarrow$ 1 hour

Lower Bound for Comparison Sorting

Goal: Determine minimum time required to sort $n$ items
Note: we want worst-case, not best-case time
- Best-case doesn’t tell us much. E.g. Insertion Sort takes $O(n)$ time on already-sorted input
- Want to know worst-case time for best possible algorithm

• How can we prove anything about the best possible algorithm?
• Want to find characteristics that are common to all sorting algorithms
• Limit attention to comparison-based algorithms and try to count number of comparisons

What you need to know / be able to do

- Know the definition of $f(n) = O(g(n))$
- Be able to prove that some function $f(n)$ is $O(g(n))$. The simplest way is as done on two slides.
- Know worst-case and average (expected) case $O(\ldots)$ of basic searching/sorting algorithms: linear/binary search, partition alg of Quicksort, insertion sort, selection sort, quicksort, merge sort.
- Be able to look at an algorithm and figure out its worst case $O(\ldots)$ based on counting basic steps or things like array-element swaps/
Comparison Trees

- Comparison-based algorithms make decisions based on comparison of data elements
- Gives a comparison tree
- If algorithm fails to terminate for some input, comparison tree is infinite
- Height of comparison tree represents worst-case number of comparisons for that algorithm
- Can show: Any correct comparison-based algorithm must make at least \( n \log n \) comparisons in the worst case

Lower Bound for Comparison Sorting

- Say we have a correct comparison-based algorithm
- Suppose we want to sort the elements in an array \( b[] \)
- Assume the elements of \( b[] \) are distinct
- Any permutation of the elements is initially possible
- When done, \( b[] \) is sorted
- But the algorithm could not have taken the same path in the comparison tree on different input permutations

Lower Bound for Comparison Sorting

How many input permutations are possible? \( n! \approx 2^n \log n \)

For a comparison-based sorting algorithm to be correct, it must have at least that many leaves in its comparison tree

To have at least \( n! \approx 2^n \log n \) leaves, it must have height at least \( n \log n \) (since it is only binary branching, the number of nodes at most doubles at every depth)

Therefore its longest path must be of length at least \( n \log n \), and that it its worst-case running time