L08: Analysis of Algorithms

- Reading: Rosen 3.1, 3.2, 3.3

Revisiting the Selection Sort Algorithm

```
(1) for i = 1 to n - 1
(2)     for j = i + 1 to n
(3)         if ( A[i] > A[j] )
(4)             swap A[i] and A[j]
(5)         endif
(6)     endfor
(7) endfor
```

Muhammad ibn Musa al-Khwarizmi

An algorithm is a finite set of precise instructions for performing a computation or for solving a problem.

How to Measure the Running Time?

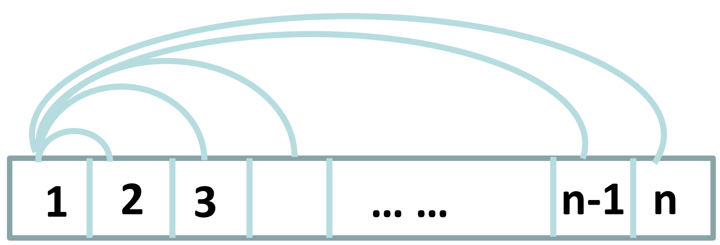
- The real running time when executed on a computer?
- The total number of lines executed?
- The total number of machine instructions, but…
  - Ignore lower order terms
  - Ignore constant coefficients
    - A constant is any quantity that doesn't depend on n.
- The number of times a particular line is executed (as a function of input size n)?
  - We will show that line (3) is executed $n(n-1)/2$ times
  - Then conclude that the running time of selection sort is $\Theta(n^2)$

Solution:

When i=1, the index j iterates from 2 to n, making a total of $n-1$ **comparisons**. Whenever the element in A[j] is smaller than in A[1], A[j] is swapped with A[1], keeping the smaller element in A[1]. After these comparisons, A[1] contains the smallest element in the array A[1,..,n].

# of comparisons

i=1:



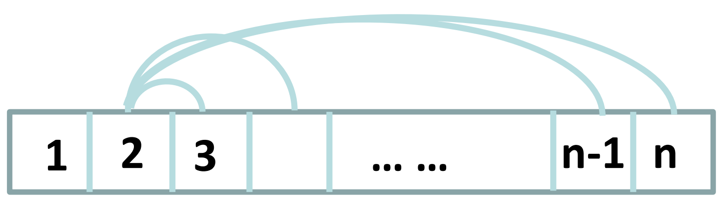
$n-1$

Solution (cont'd):

When i=2, index j iterates from 3 to n (i.e. **n-2 comparisons**), effectively comparing all the elements in A[2,…,n] and resulting in the smallest element of A[2,..,n] being kept in A[2]. Therefore A[2] contains the second smallest element in A[1,…,n] (while A[1] contains the smallest).

# of comparisons

i=2:



$n-2$

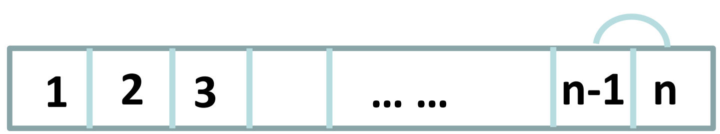
Solution (cont'd):

When i=n-1, j iterates once, from n to n (i.e. **1 comparison**).

A[n-1] will contain the $(n\text{-}1)^{th}$ smallest number of the array A[1,…,n].

\# of comparisons

i=n-1:

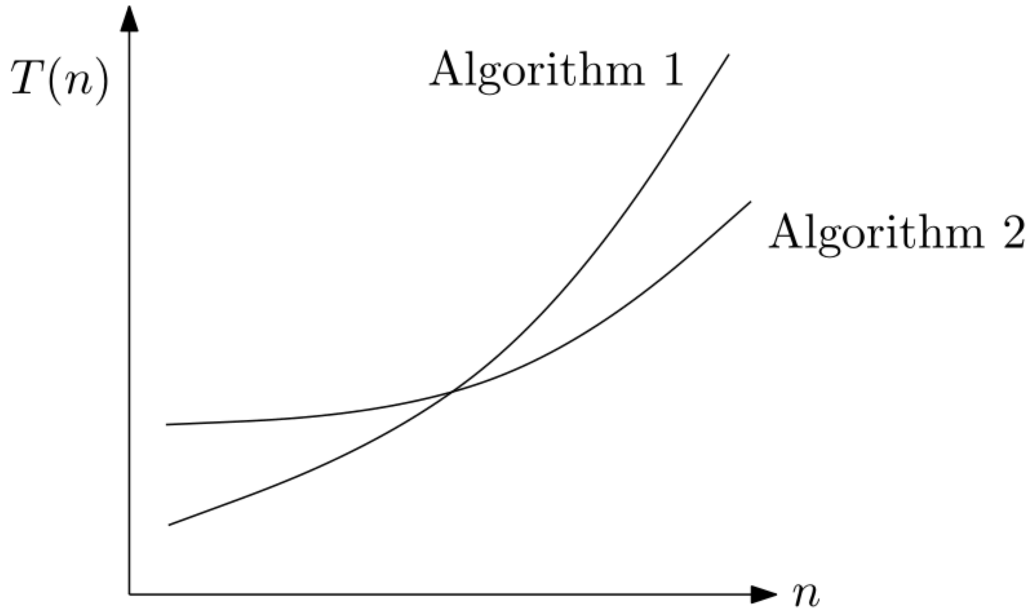


1

- Therefore total # of comparisons = $(n-1)+(n-2)+\cdots+1$

  $= \boldsymbol{n(n-1)/2}$

The Growth of Functions



- Which algorithm is better for large n?
  - For Algorithm 1, $T_1(n) = 3n^3 + 6n^2 - 4n + 17 = \Theta(n^3)$
  - For Algorithm 2, $T_2(n) = 7n^2 - 8n + 20 = \Theta(n^2)$
  - Clearly, Algorithm 2 is better

Big-Theta

- **Definition**: Let f and g be functions from the set of positive real numbers to the set of positive real numbers. We say that $f(x)$ is $\Theta(g)$ if there are positive constants C_1, C_2, and k such that
$$C_1 g(x) \leq f(x) \leq C_2 g(x)$$
whenever $x > k$.
- This is read as "$f(x)$ is big-Theta of $g(x)$" or "$f(x)$ is asymptotically the same as $g(x)$."
- Usually written as $f(n) = \Theta\big(g(n)\big)$, although the more mathematically correct way should be $f(n) \in \Theta\big(g(n)\big)$.
- The constants C_1, C_2 and k are called *witnesses* to the relationship. There are infinitely many such witnesses. Only one pair of witnesses is needed for lower/upper bound.

Using Definition to Derive Big-Theta

$$T_1(n) = 3n^3 + 6n^2 - 4n + 17 = \Theta(n^3)$$

- Choose $C_1 = 2, C_2 = 4$
- Want a k such that, when $n > k$

$$2n^3 \leq 3n^3 + 6n^2 - 4n + 17 \leq 4n^3$$
$$-n^3 \leq 6n^2 - 4n + 17 \leq n^3$$

- It's clear that such a k must exist, there is no need to actually find it.

Comparison of Algorithms

- n is big (big data!), so we are interested in

$$\lim_{n\to\infty} \frac{T_1(n)}{T_2(n)}$$

- Three cases:
  - $\lim_{n\to\infty} \frac{T_1(n)}{T_2(n)} = 0$: Algorithm 1 is better
  - $\lim_{n\to\infty} \frac{T_1(n)}{T_2(n)} = \infty$: Algorithm 2 is better
  - $\lim_{n\to\infty} \frac{T_1(n)}{T_2(n)} = C$ for some constant $0 < C < \infty$, or $\frac{T_1(n)}{T_2(n)}$ oscillates: Θ-notation cannot tell, need more careful analysis.
- If $T_1(n) = \Theta\big(g_1(n)\big), T_2(n) = \Theta\big(g_2(n)\big)$, it's sufficient to consider $\lim_{n\to\infty} \frac{g_1(n)}{g_2(n)}$

Examples

- $\log_{10} n = \frac{\log_2 n}{\log_2 10} = \Theta(\log_2 n) = \Theta(\log n)$
- $9999^{9999^{9999}} = \Theta(1)$
- 2^{10n} is not $\Theta(2^n)$, 3^n is not $\Theta(2^n)$
- $\sum_{i=1}^{n} i^2 \leq n^2 \cdot n \leq n^3$
  $\sum_{i=1}^{n} i^2 \geq \left(\frac{n}{2}\right)^2 \cdot \left(\frac{n}{2}\right) \geq \frac{1}{8} n^3$
  So, $\sum_{i=1}^{n} i^2 = \Theta\left(n^3\right)$
- $\sum_{i=0}^{n} c^i = \frac{c^{n+1}-1}{c-1} = \begin{cases} \Theta(c^n), c > 1 \\ \Theta(n), c = 1 \\ \Theta(1), c < 1 \end{cases}$ (geometric series)

Solving Geometric Series

- Geometric series

$$S(n) = 1 + c + c^2 + c^3 + \cdots + c^n$$

- Solving geometric series

$$S(n+1) = S(n) \cdot c + 1$$
$$S(n+1) = S(n) + c^{n+1}$$
$$S(n) \cdot c + 1 = S(n) + c^{n+1}$$
$$(c-1)S(n) = c^{n+1} - 1$$

- If $c \neq 1$,

$$S(n) = \frac{c^{n+1} - 1}{c - 1}$$

Examples

- $\log(n!) = \log(n) + \log(n-1) + \cdots + \log 1 \le n \log n$
  $\log(n!) \ge \log(n) + \log(n-1) + \cdots + \log\left(\frac{n}{2}\right) \ge \frac{n}{2}\log\left(\frac{n}{2}\right)$
  So, $\log(n!) = \Theta(n \log n)$.
- $\sum_{i=1}^{n} \frac{1}{i} = \Theta(\log n)$ (harmonic series, derivation on board)

Examples

- c_1^n dominates n^{c_2}, which dominates $\log^{c_3} n$, for $c_1 > 1, c_2 > 0, c_3 > 0$
- $n^{0.1} + \log^{10} n = \Theta\left(n^{0.1}\right)$
- $1.1^n + n^{100} = \Theta(1.1^n)$

Limitation of Big-Theta

- Some functions cannot be described by Θ
  - Example:
    the number of 1's in the binary representation of n
  - Oscillates between 1 and $\log n$
- Some functions are hard to describe by Θ
  - Example: $n!$
  - It is known that $n! = \Theta\left(\sqrt{n}\left(\frac{n}{e}\right)^n\right)$ (Stirling's formula)
    but it's very difficult to derive
- When used for analysis of algorithms (later)

Big-Oh

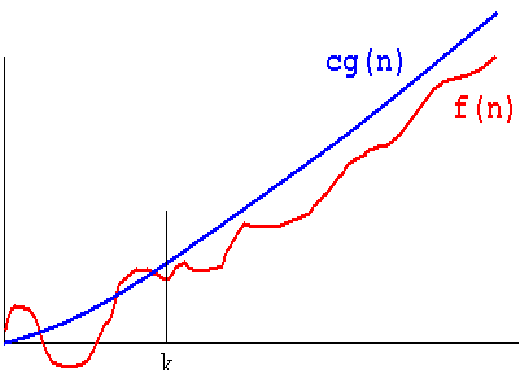
- **Definition**: Let f and g be functions from the set of positive real numbers to the set of positive real numbers. We say that $f(x)$ is $O(g)$ if there are positive constants C_2, and k such that
$$f(x) \leq C_2 g(x)$$
whenever $x > k$.
- This is read as "$f(x)$ is big-Oh of $g(x)$" or "$f(x)$ is asymptotically dominated by $g(x)$."
- Usually written as $f(n) = O\big(g(n)\big)$, although the more mathematically correct way should be $f(n) \in O\big(g(n)\big)$.
- The constants C_2 and k are called *witnesses* to the relationship. Only one pair of witnesses is needed.

Big-Omega

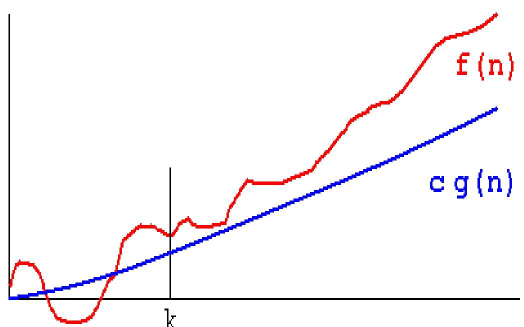
- **Definition**: Let f and g be functions from the set of positive real numbers to the set of positive real numbers. We say that $f(x)$ is $\Omega(g)$ if there are positive constants C_1, and k such that
$$C_1 g(x) \leq f(x)$$
whenever $x > k$.
- This is read as "$f(x)$ is big-Omega of $g(x)$" or "$f(x)$ asymptotically dominates $g(x)$."
- Usually written as $f(n) = \Omega\big(g(n)\big)$, although the more mathematically correct way should be $f(n) \in \Omega\big(g(n)\big)$.
- The constants C_1 and k are called *witnesses* to the relationship. Only one pair of witnesses is needed.
- $f(x) = \Theta\big(g(x)\big)$ iff $f(x) = O\big(g(x)\big)$ and $f(x) = \Omega\big(g(x)\big)$

Review

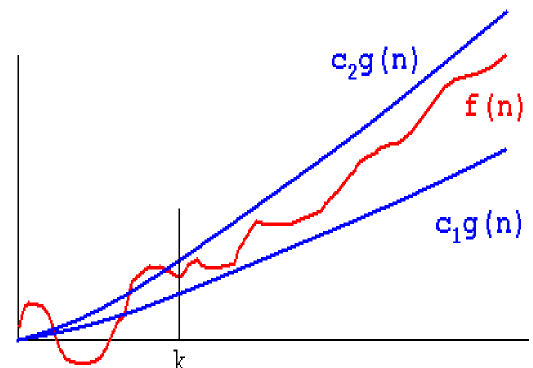
If $f(n)$ is:	Then <u>for large enough</u> n it is:	We say that:
$O(g(n))$	Upper bounded by $c \cdot g(n)$	"$f(n)$ is dominated by $g(n)$"
$\Omega(g(n))$	Lower bounded by $c \cdot g(n)$	"$f(n)$ dominates $g(n)$"
$\Theta(g(n))$	Lower bounded by $c_1 \cdot g(n)$ and upper bounded by $c_2 \cdot g(n)$	"$f(n)$ grows asymptotically with $g(n)$"



$O(g(n))$



$\Omega(g(n))$



$\Theta(g(n))$

$$\boldsymbol{f(x) = \Theta\big(g(x)\big) \text{ iff } f(x) = O\big(g(x)\big) \text{ and } f(x) = \Omega\big(g(x)\big)}$$

Examples:

- $f(n) = 32n^2 + 17n - 32$.
  - $f(n)$ is $O(n^2), O(n^3), \Omega(n^2), \Omega(n)$, and $\Theta(n^2)$.
  - $f(n)$ is not $O(n), \Omega(n^3), \Theta(n)$, or $\Theta(n^3)$.
- The number of 1's in the binary representation of n is
  - $O(\log n)$
  - $\Omega(1)$
- $n!$
  - $n! \leq n^n = O(n^n)$
  - $n! \geq \left(\frac{n}{2}\right)^{\frac{n}{2}} = \Omega\left(\left(\frac{n}{2}\right)^{\frac{n}{2}}\right)$

Insertion Sort

<u>Insertion-Sort(A):</u>
for $j \leftarrow 2$ **to** n **do**
    $key \leftarrow A[j]$
    $i \leftarrow j - 1$
    while $i \geq 1$ **and** $A[i] > key$ **do**
        $A[i+1] \leftarrow A[i]$
        $i \leftarrow i - 1$
    endwhile
    $A[i+1] \leftarrow key$
endfor

sorted	key	unsorted

Insertion Sort: Example

- 1st iteration:
  - (4 1 8 2 5) → (4 4 8 2 5) → (1 4 8 2 5)
  - key = 1
- 2nd iteration:
  - (1 4 8 2 5)
  - key = 8
- 3rd iteration:
  - (1 4 8 2 5) → (1 4 8 8 5) → (1 4 4 8 5) → (1 2 4 8 5)
  - key = 2
- 4th iteration:
  - (1 2 4 8 5) → (1 2 4 8 8) → (1 2 4 5 8)
  - key = 5

Analysis of Algorithms

- Question: What's the running time of insertion sort if the input array is already sorted?
- Answer: $\Theta(n)$.
- Question: What's the running time of insertion sort if the input array is inversely sorted?
- Answer: $\Theta(n^2)$.
- Observation: The running time of an algorithm doesn't only depend on n, it also depends on the actual input!
- Question: Which input should we use for analyzing an algorithm?
  - Best-case? Average-case? Worst-case?

Worst-case Analysis

- The default of algorithm analysis
- Especially easy when using big-Oh
  - You don't really have to find the worst-case input!
  - Can easily conclude that insertion sort runs in time $O(n^2)$.
- How to show an algorithm has worst-case running time $\Theta(n^2)$?
  - Show that it's $O(n^2)$
  - Show that it's $\Omega(n^2)$
    - Find an input such that the running time is $\Omega(n^2)$

Example: Linear Search

- Problem: Given an array A of unordered elements and x, find x or report that x doesn't exist in A
- Algorithm:

procedure *linear search*(x:integer,
$a_1, a_2, \ldots, a_n$: distinct integers)
$i := 1$
while ($i \leq n$ and $x \neq a_i$)
    $i := i + 1$
if $i \leq n$ **then** *location* $:= i$
else *location* $:= 0$
return *location*

Analysis of Linear Search

- Running time is $O(n)$
  - Obvious, since the loop has at most n iterations.
- (Worst-case) running time is $\Omega(n)$
  - When x doesn't exist in A, the loop has exactly n iterations.
- Running time is $\Theta(n)$

Example: Binary Search

- Problem: Given an array A of ordered elements and x, find x, or report that x doesn't exist in A
- Algorithm:

procedure binary search(x: integer, $a_1, a_2, \ldots, a_n$: increasing integers)
$i := 1$ {i is the left endpoint of interval}
$j := n$ {j is right endpoint of interval}
while $i <= j$
    $m := \lfloor (i + j)/2 \rfloor$
    if $x = a_m$ then **return** m
    if $x > a_m$ then $i :=$ m + 1
    else $j :=$ m – 1
return -1 (x doesn't exist)

Analysis of Binary Search

- Assumption: $n = 2^k$
- Running time is $O(k) = O(\log n)$
  - The length of the range $j - i + 1$ decrease by half in each iteration
  - The algorithm terminates when $i \geq j$, i.e.,
$$j - i + 1 \leq 1$$
- (Worst-case) running time is $\Omega(\log n)$
  - When x doesn't exist in A, the loop has exactly $\log n$ iterations.
- Running time is $\Theta(\log n)$
- What if n is not in the form of 2^k?
  - Find k such that $2^k < n < 2^{k+1}$. The running time must be between $\Theta(k)$ and $\Theta(k+1)$, which is $\Theta(k)$.