RETURN OF INSERTION-SORT

```
INSERTION-SORT(A)                          (input in array A)
1   for j := 2 to A.length do              (move the limit of the sorted range)
2       key := A[j]                        (handle the first unsorted element)
3       i := j − 1
4       while i > 0 and A[i] > key do      (find the correct location of the new element)
5           A[i + 1] := A[i]               (make room for the new element)
6           i := i − 1
7       A[i + 1] := key                    (set the new element to it's correct location)
```

- line 1 is executed n times
- lines 2 and 3 are executed n - 1 times
- line 4 is executed at least n - 1 and at most (2 + 3 + 4 + $\cdots$ + n - 2) times
- lines 5 and 6 are executed at least 0 and at most (1 + 2 + 3 + 4 + $\cdots$ + n - 3) times

- in the best case the entire array is already sorted and the running time of the entire algorithm is at least $\Theta(n)$
- in the worst case the array is in a reversed order. $\Theta(n^2)$ time is used
- once again determining the average case is more difficult:
- let's assume that out of randomly selected element pairs half is in an incorrect order in the array

  $\Rightarrow$ the amount of comparisons needed is half the amount of the worst case where all the element pairs were in an incorrect order

  $\Rightarrow$ the average-case running time is the worst-case running time divided by two: ((n - 1)n)/ 4 = $\Theta(n^2)$

4.2 Algorithm Design Technique: Divide and Conquer

We've earlier seen the *decrease and conquer* algorithm design technique and the algorithm INSERTION-SORT as an example of it.

Now another technique called *divide and conquer* is introduced. It is often more efficient than the decrease and conquer approach.

- the problem is divided into several subproblems that are like the original but smaller in size.
- small subproblems are solved straightforwardly
- larger subproblems are further divided into smaller units
- finally the solutions of the subproblems are combined to get the solution to the original problem

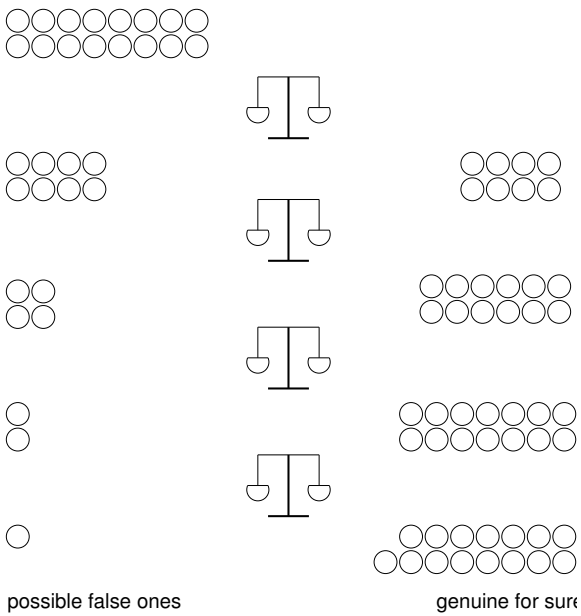
Let's get back to the claim made earlier about the complexity notation not being fixed to programs and take an everyday, concrete example

Example: finding the false goldcoin

- The problem is well-known from logic problems.
- We have n gold coins, one of which is false. The false coin looks the same as the real ones but is lighter than the others. We have a scale we can use and our task is to find the false coin.
- We can solve the problem with Decrease and conquer by choosing a random coin and by comparing it to the other coins one at a time.

  $\Rightarrow$ At least 1 and at most n - 1 weighings are needed. The best-case efficiency is $\Theta(1)$ and the worst and average case efficiencies are $\Theta(n)$.
- Alternatively we can always take two coins at random and weigh them. At most $n/2$ weighings are needed and the efficiency of the solution is still the same.

The same problem can be solved more efficiently with divide and conquer:



- Divide the coins into the two pans on the scales. The coins on the heavier side are all authentic, so they don't need to be investigated further.
- Continue the search similarly with the lighter half, i.e. the half that contains the false coin, until there is only one coin in the pan, the coin that we know is false.
- The solution is recursive and the base case is the situation where there is only one possible coin that can be false.

- The amount of coins on each weighing is 2 to the power of the amount of weighings still required: on the highest level there are $2^{weighings}$ coins, so based on the definition of the logarithm:

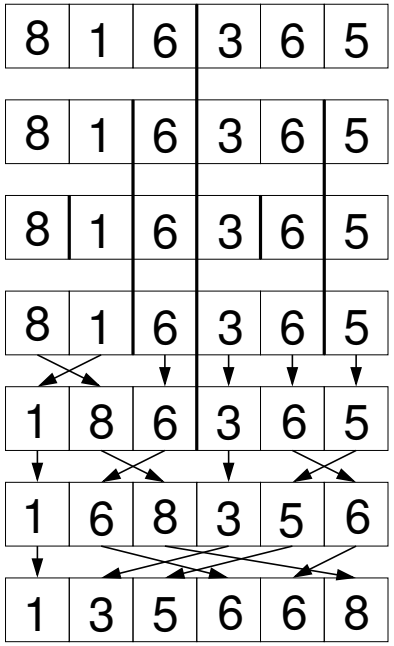
$$2^{weighings} = n \Rightarrow log_2 n = weighings$$

- Only $\log_2 n$ weighings is needed, which is significantly fewer than $n/2$ when the amount of coins is large.

  $\Rightarrow$ The complexity of the solution is $\Theta(\lg n)$ both in the best and the worst-case.

MERGE-SORT:

- divide the elements in the array into two halves.
- continue dividing the halves further in half until the subarrays contain atmost one element
- arrays of 0 or 1 length are already sorted and require no actions
- finally merge the sorted subarrays



- the MERGE-algorithm for merging the subarrays:

```
MERGE(A, left, middle, right)
1   for i := left to right do      (scan through the entire array...)
2       B[i] := A[i]               (... and copy it into a temporary array)
3   i := left                      (set i to indicate the endpoint of the sorted part)
4   j := left; k := middle + 1     (set j and k to indicate the beginning of the subarrays)
5   while j ≤ middle and k ≤ right do      (scan until either half ends)
6       if B[j] ≤ B[k] then        (if the first element in the lower half is smaller...)
7           A[i] := B[j]           (... copy it into the result array...)
8           j := j + 1             (... increment the starting point of the lower half)
9       else                       (else...)
10          A[i] := B[k]           (... copy the first element of the upper half...)
11          k := k + 1             (... and increment its starting point)
12      i := i + 1                 (increment the starting point of the finished set)
13  if j > middle then
14      k := 0
15  else
16      k := middle − right
17  for j := i to right do         (copy the remaining elements to the end of the result)
18      A[j] := B[j + k]
```

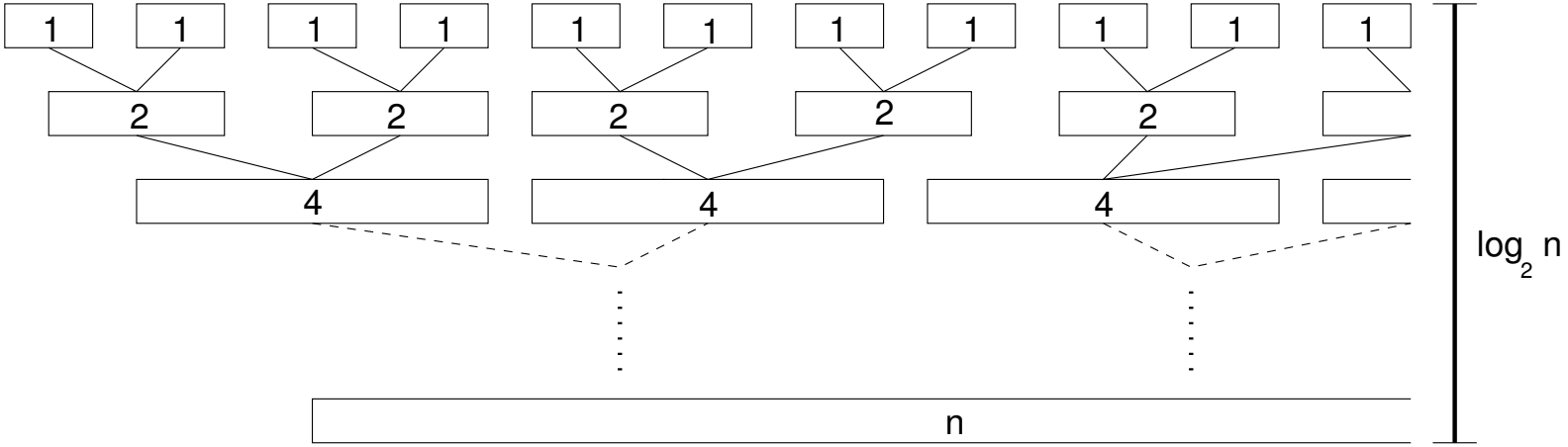
- MERGE-SORT

```
MERGE-SORT(A, left, right)
1  if left < right then                    (if there are elements in the array...)
2      middle := ⌊(left + right)/2⌋        (... divide it into half)
3      MERGE-SORT(A, left, middle)         (sort the upper half...)
4      MERGE-SORT(A, middle + 1, right)    (... and the lower)
5      MERGE(A, left, middle, right)       (merge the parts maintaining the order)
```

- MERGE merges the arrays by using the "decrease and conquer" method.
  - the first **for**-loop uses a linear amount of time $\Theta(n)$ relative to the size of the subarray
  - the **while**-loop scans through the upper and the lower halves at most once and at least one of the halves entirely $\Rightarrow \Theta(n)$
  - the second **for**-loop scans through at most half of the array and its running time is $\Theta(n)$ in the worst case.
  - other operations are constant time

$\Rightarrow$ the running time of the entire algorithm is obtained by combining the results. It's $\Theta(n)$ both in the best and the worst case.

- Like with QUICKSORT, the analysis of MERGE-SORT is not as straghtforward since it is recursive.
- MERGE-SORT calls itself and MERGE, all other operations are constant time $\Rightarrow$ we can concentrate on the time used by the intances of MERGE, everything else is constant time.



- The instances of MERGE form a treelike structure shown on the previous page.
  - the sizes of the subarrays are marked on the instances of MERGE in the picture
  - on the first level the length of each subarray is 1 (or 0)
  - the subarrays on the upper levels are always two times as large as the ones on the previous level
  - the last level handles the entire array
  - the combined size of the subarrays on each level is n
  - the amount of instances of MERGE on a given level is two times the amount on the previous level.
    $\Rightarrow$ the amount increases in powers of two, so the amount of instances on the last level is 2^h, where h is the height of the tree.
  - on the last level there are approximately n instances
    $\Rightarrow 2^h = n \Leftrightarrow \log_2 n = h$, the height of the tree is $\log_2 n$
  - since a linear amount of work is done on each level and there are $\lg n$ levels, the running time of the entire algorithm is $\Theta(n \lg n)$

MERGE-SORT is clearly more complicated than INSERTION-SORT. Is using it really worth the effort?

Yes, on large inputs the difference is clear.

- if n is 1000 000 n^2 is 1000 000 000 000, while n log n is about 19 930 000

Advantages and disadvantages of Mergesort

Advantages

- Running time $\Theta(n \lg n)$
- Stable

Disadvantages

- MERGE-SORT requires $\Theta(n)$ extra memory, INSERTION-SORT and QUICKSORT sort in place.
- Constant coefficient quite large

4.3 Return of Quicksort

Let's next cover a very efficient sorting algorithm QUICKSORT.

QUICKSORT is a divide and conquer algorithm.

The division of the problem into smaller subproblems

- Select one of the elements in the array as a *pivot*, i.e. the element which partitions the array.
- Change the order of the elements in the array so that all elements smaller or equal to the pivot are placed before it and the larger elements after it.
- Continue dividing the upper and lower halves into smaller subarrays, until the subarrays contain 0 or 1 elements.

Smaller subproblems:

- Subarrays of the size 0 and 1 are already sorted

Combining the sorted subarrays:

- The entire (sub) array is automatically sorted when its upper and lower halves are sorted.
  - all elements in the lower half are smaller than the elements in the upper half, as they should be

QUICKSORT-algorithm

QUICKSORT$(A, left, right)$

```
1  if left < right then                          (do nothing in the trivial case)
2      pivot := PARTITION(A, left, right)        (partition in two)
3      QUICKSORT(A, left, pivot − 1)             (sort the elements smaller than the pivot)
4      QUICKSORT(A, pivot + 1, right)            (sort the elements larger than the pivot)
```

The *partition algorithm* rearranges the subarray in place

```
PARTITION(A, left, right)
1  pivot := A[right]                    (choose the last element as the pivot)
2  i := left − 1                        (use i to mark the end of the smaller elements)
3  for j := left to right − 1 do        (scan to the second to last element)
4      if A[j] ≤ pivot                  (if A[j] goes to the half with the smaller elements...)
5          i := i + 1                   (... increment the amount of the smaller elements...)
6          exchange A[i] ↔ A[j]         (... and move A[j] there)
7  exchange A[i + 1] ↔ A[right]         (place the pivot between the halves)
8  return i + 1                         (return the location of the pivot)
```

How fast is PARTITION?

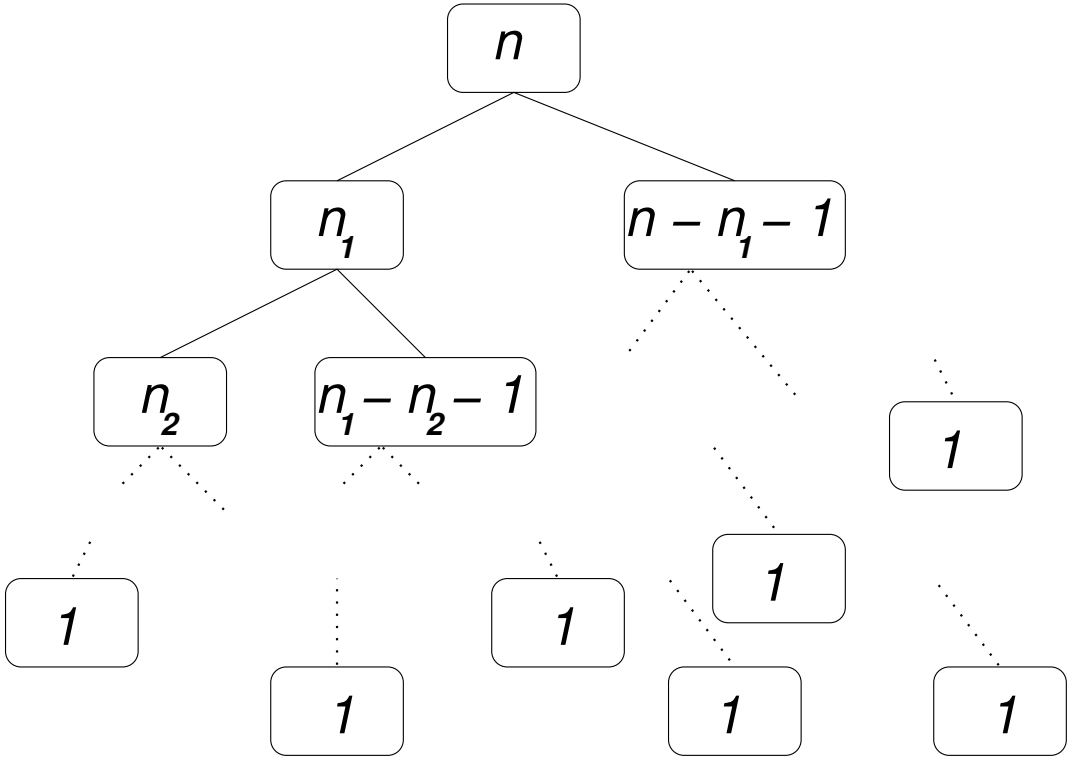
- The **for**-loop is executed n - 1 times when n is r - p
- All other operations are constant time.

$\Rightarrow$ The running-time is $\Theta(n)$.

Determining the running-time of QUICKSORT is more difficult since it is recursive. Therefore the equation for its running time would also be recursives.

Finding the recursive equation is, however, beyond the goals of this course so we'll settle for a less formal approach

- As all the operations of QUICKSORT except PARTITION and the recursive call are constant time, let's concentrate on the time used by the instances of PARTITION.



- The total time is the sum of the running times of the nodes in the picture abowe.
- The execution is constant time for an array of size 1.
- For the other the execution is linear to the size of the array.
  $\Rightarrow$ The total time is Θ(the sum of the numbers of the nodes).

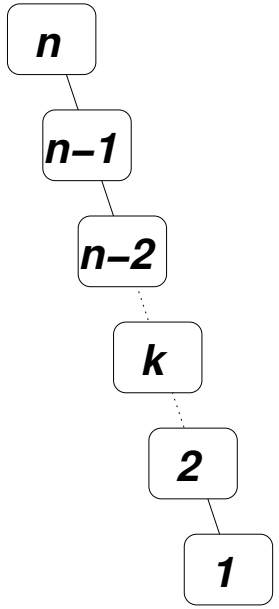
Worst-case running time

- The number of a node is always smaller than the number of its parent, since the pivot is already in its correct location and doesn't go into either of the sorted subarrays

  $\Rightarrow$ there can be atmost n levels in the tree

- the worst case is realized when the smallest or the largest element is always chosen as the pivot
  - this happens, for example, with an array already sorted
- the sum of the node numbers is n + n - 1 + $\cdots$ + 2 + 1

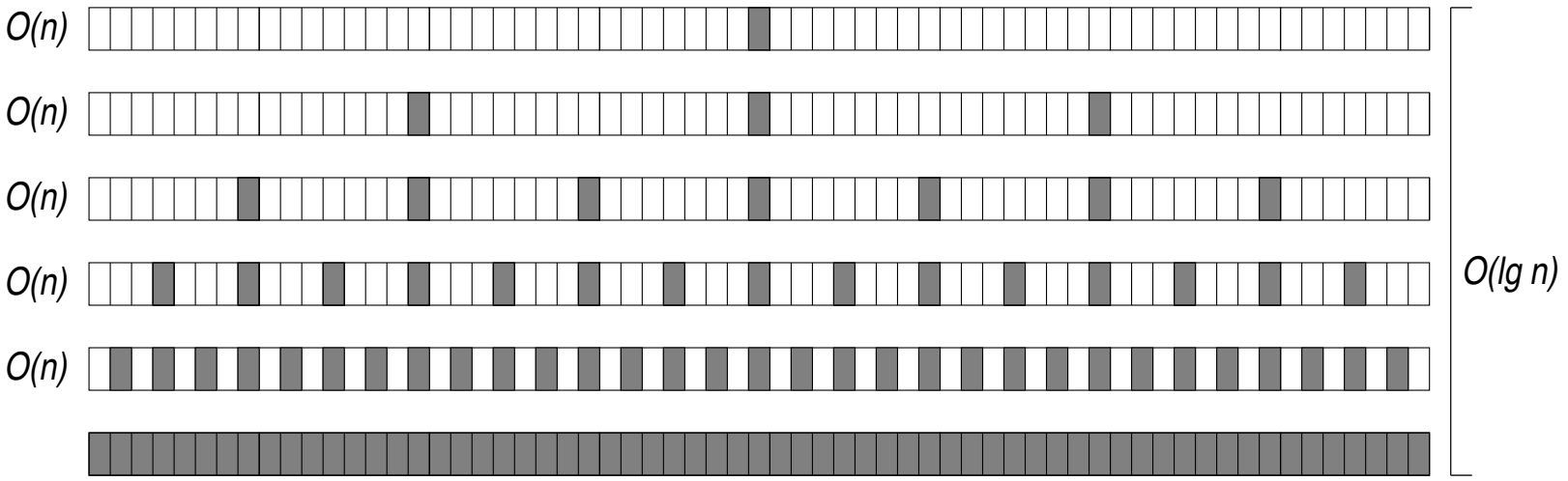
$\Rightarrow$ the worst case running time of QUICKSORT is $\Theta(n^2)$



The best-case is when the array is always divided evenly in half.

- The picture below shows how the subarrays get smaller.
  - The grey boxes mark elements already in their correct position.
- The amount of work on each level is in $\Theta(n)$.
  - a pessimistic estimate on the height of the execution tree is in the best-case $\Rightarrow \Theta(\lg n)$

$\Rightarrow$ The upper limit for the best-case efficiency is $\Theta(n \lg n)$.



The best-case and the worst-case efficiencies of QUICKSORT differ significantly.

- It would be interesting to know the average-case running-time.
- Analyzing it is beyond the goals of the course but it has been shown that if the data is evenly distributed its average running-time is $\Theta(n \lg n)$.
- Thus the average running-time is quite good.

An unfortunate fact with QUICKSORT is that its worst-case efficiency is poor and in practise the worst-case situation is quite probable.

- It is easy to see that there can be situations where the data is already sorted or almost sorted.

$\Rightarrow$ A way to decrease the risk of the systematic occurence of the worst-case situation's likelyhood is needed.

Randomization has proved to be quite efficient.

Advantages and disadvantages of QUICKSORT

Advantages:

- sorts the array very efficiently in average
  - the average-case running-time is $\Theta(n \lg n)$
  - the constant coefficient is small
- requires only a constant amount of extra memory
- if well-suited for the virtual memory environment

Disadvantages:

- the worst-case running-time is $\Theta(n^2)$
- without randomization the worst-case input is far too common
- the algorithm is recursive
  $\Rightarrow$ the stack uses extra memory
- instability

5 Complexity notations

This chapter discusses the notations used to describe the asymptotic behaviour of algorithms.

Θ is defined together with two other related useful notations O and Ω.

5.1 Asymptotic notations

The equation for the running time was simplified earlier significantly:

- only the highest order term was used
- its constant coefficient was left out

$\Rightarrow$ studying the behaviour of the algorithm as the size of its input increases to infinity

- i.e. the *asymptotic* efficiency of algorithms

$\Rightarrow$ usefull information **only with inputs larger than a certain limit**

- often the limit is rather low

  $\Rightarrow$ the algorithm fastest according to Θ- and other notations is the fastest also in practice, except on very small inputs

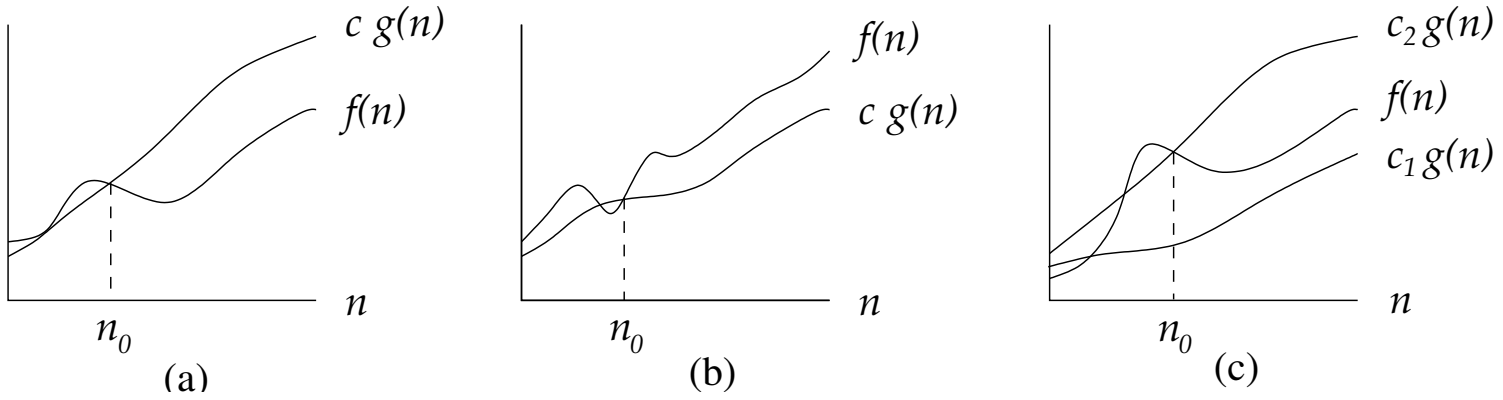
Θ-notation

- let $g(n)$ be a function from a set of numbers to a set of numbers

  $\Theta(g(n))$ is the set of those functions $f(n)$ for which there exists positive constants c_1, c_2 and n_0 so that for all $n \geq n_0$

  $$0 \leq c_1 \cdot g(n) \leq f(n) \leq c_2 \cdot g(n)$$

- the function in the picture (c) $f(x) = \Theta(g(n))$
  - $\Theta(g(n))$ is a *set* of functions
    $\Rightarrow$ we should, and earlier did, write $f(n) \in \Theta(g(n))$, but usually $=$ is used.



Whether a function $f(n)$ is $\Theta(g(n))$ can be proven by finding suitable values for the constants c_1, c_2 and n_0 and by showing that the function stays larger or equal to $c_1 g(n)$ and smaller or equal to $c_2 g(n)$ with values of n starting from n_0.

For example: $3n^2 + 5n - 20 = \Theta(n^2)$

- let's choose $c_1 = 3$, $c_2 = 4$ ja $n_0 = 4$
- $0 \leq 3n^2 \leq 3n^2 + 5n - 20 \leq 4n^2$ when $n \geq 4$, since $0 \leq 5n - 20 \leq n^2$
- just as well we could have chosen $c_1 = 2$, $c_2 = 6$ and $n_0 = 7$ or $c_1 = 0{,}000\,1$, $c_2 = 1\,000$ and $n_0 = 1\,000$
- what counts is being able to choose *some* positive c_1, c_2 and n_0 that fullfill the requirements

An important result: if $a_k > 0$, then

$$a_k n^k + a_{k-1} n^{k-1} + \cdots + a_2 n^2 + a_1 n + a_0 = \Theta(n^k)$$

- in other words, if the coefficient of the highest-order term of a polynomial is positive the Θ-notation allows us to ignore all lower-order terms and the coefficient.

The following holds for constant functions $c = \Theta(n^0) = \Theta(1)$

- $\Theta(1)$ doesn't indicate which variable is used in the analysis
  $\Rightarrow$ it can only be used if the variable is clear from the context

O-notation (pronounced "big-oh")

The O-notation is otherwise like the Θ-notation, but it bounds the function only from above.
$\Rightarrow$ *asymptotic upper bound*

Definition:

> *$O(g(n))$ is the set of functions $f(n)$ for which there exists positive constants c and n_0 such that, for all $n \geq n_0$*
>
> $$0 \leq f(n) \leq c \cdot g(n)$$

- the function in the picture (a) $f(x) = O(g(n))$
- it holds: if $f(n) = \Theta(g(n))$, then $f(n) = O(g(n))$
- the opposite doesn't always hold: $n^2 = O(n^3)$, but $n^2 \neq \Theta(n^3)$
- an important result: if $k \leq m$, then $n^k = O(n^m)$
- if the running time of the **slowest** case is $O(g(n))$, then the running time of **every** case is $O(g(n))$

The O-notation is important in practise as, for example, the running times quaranteed by the C++-standard are often given in it.

Often some upper bound can be given to the running time of the algorithm in the slowest possible case with the O-notation (and every case at the same time).

We are often interested in the upper bound only.

For example: INSERTION-SORT

line	efficiency
for j := 2 **to** $A.length$ **do**	$O(n)$
key := $A[j]$	$\cdot \quad O(1)$
i := $j-1$	$\cdot \quad O(1)$
while $i > 0$ and $A[i] > key$ **do**	$\cdot \quad O(n)$
$A[i+1]$:= $A[i]$	$\cdot \quad \cdot \quad O(1)$
i := $i-1$	$\cdot \quad \cdot \quad O(1)$
$A[i+1]$:= key	$\cdot \quad O(1)$

The worst case running time is $O(n) \cdot O(n) \cdot O(1) = O(n^2)$

Ω-notation (pronounced "big-omega")

The Ω-notation is otherwise like the Θ-notation but is bounds the function only from below.
$\Rightarrow$ *asymptotic lower bound*

Definition:

> $\Omega(g(n))$ *is the set of functions* $f(n)$ *for which there exist positive constants* c *and* n_0 *such that, for all* $n \geq n_0$
>
> $$0 \leq c \cdot g(n) \leq f(n)$$

- the function in the picture (b) function $f(x) = \Omega(g(n))$
- the following result follows from the definitions: $f(n) = \Theta(g(n))$ *if and only if* $f(n) = O(g(n))$ *and* $f(n) = \Omega(g(n))$.
- if the running time of the **fastest** case is $\Omega(g(n))$, the running time of **every** case is $\Omega(g(n))$

The Ω-notation is usefull in practise mostly in situations where even the best-case efficiency of the solution is unsatisfactory and the result can be rejected straightaway

Properties of asymptotic notations

$f(n) = \Omega(g(n))$ and $f(n) = O(g(n)) \iff f(n) = \Theta(g(n))$

Many of the relational properties of real numbers apply to asymptotic notations:

$$\begin{aligned} f(n) &= O(g(n)) \quad a \leq b \\ f(n) &= \Theta(g(n)) \quad a = b \\ f(n) &= \Omega(g(n)) \quad a \geq b \end{aligned}$$

i.e. if the highest-order term of $f(n)$ whose constant coefficient has been removed $\leq g(n)$'s corresponding term, $f(n) = O(g(n))$ etc.

Note the difference: for any two real numbers exactly one of the following must hold: $a < b$, $a = b$ ja $a > b$. However, this does not hold for all asymptotic notations.
$\Rightarrow$ Not all functions are asymptotically comparable to eachother (e.g. n and $n^{1+\sin n}$).

An example simplifying things a little:

- If an algorithm is $\Omega(g(n))$, its consumption of resources is at least $g(n)$.
  - cmp. a book costs at least about 10 euros.
- If an algorithm is $O(g(n))$, its consumption of resources is at most $g(n)$.
  - cmp. a book costs at most 10 euros.
- If an algorithms is $\Theta(g(n))$, its consumption of resources is always $g(n)$.
  - cmp. a book costs about 10 euros

Note that the running time of all algorithms cannot be determined with the Θ-notation.

For example Insertion-Sort:

- the best-case is $\Omega(n)$, but not $\Omega(n^2)$
- the worst-case is $O(n^2)$, but not $O(n)$
  $\Rightarrow$ a Θ-value common to all cases cannot be determined.

An example

Let's take a function $f(n) = 3n^2 + 5n + 2$.

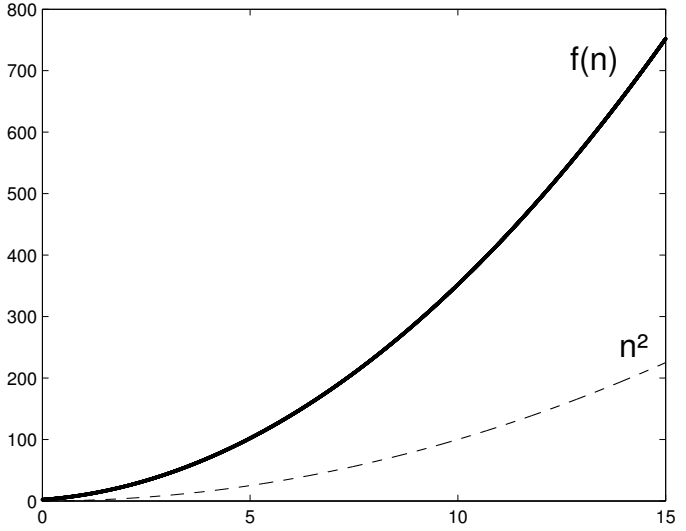
and simplify it according to the rules given earlier:

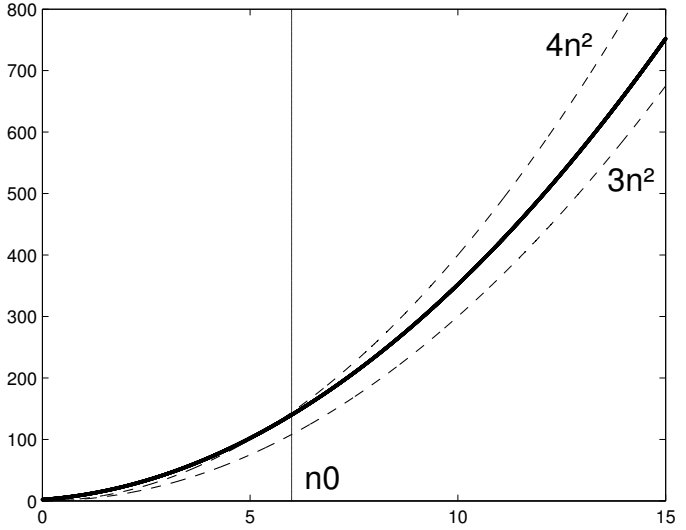
- lower-order terms ignored
- constant coefficients ignored

$\Rightarrow f(n) = \Theta(n^2)$

To be completely convinced we'll determine the constants c_1 ja c_2:

$3n^2 \leq 3n^2 + 5n + 2 \leq 4n^2$, when $n \geq 6$
$\Rightarrow c_1 = 3$, $c_2 = 4$ and $n_0 = 6$ work correctly
$\Rightarrow f(n) = O(n^2)$ and $\Omega(n^2)$
$\Rightarrow f(n) = \Theta(n^2)$





Clearly the constant $c_2 = 4$ works also when $g(n) = n^3$, since when $n \geq 6, n^3 > n^2$
$\Rightarrow f(n) = O(n^3)$

- the same holds when $g(n) = n^4$...

And below the constant $c_1 = 3$ works also when $g(n) = n \lg n$, since when $n \geq 6$, $n^2 > n \lg n$
$\Rightarrow f(n) = \Omega(n \lg n)$

- the same holds when $g(n) = n$ or $g(n) = \lg n$