The Complexity of Algorithms and the Lower Bounds of Problems

Outlines

- The Time-Complexity of an Algorithm
- The Best, Average and Worst Case Analysis of Algorithms
- The Lower Bound of a Problem
- The Worst Case Lower Bound of Sorting
- Heapsort - A Sorting Algorithm which Is Optimal in Worst Cases
- The Average Case Lower Bound of Sorting
- The Improving of a Lower Bound through Oracles
- The Finding of Lower Bound by Problem Transformation
- Linear Time Sorting Algorithm

學習目標

- 各排序演算法的設計與時間複雜度分析。
  - 插入排序法 (insertion sort)
  - 選擇排序法 (selection sort)
  - 快速排序法 (quicksort)
  - 淘汰排序法 (knockout sort)
  - 堆積排序法 (heap sort)
- Binary search演算法的設計與時間複雜度分析。
- 2-D rank finding problem演算法的設計與時間複雜度分析。

學習目標

- 以比較為基礎的排序演算法的最壞情形下時間複雜度之下界(lower bound)分析。
- 以比較為基礎的排序演算法的平均情形下時間複雜度之下界(lower bound)分析。
- 利用oracle 技術改進之下界(lower bound)分析。
- 利用reduction技術證明問題的下界(lower bound)。
- 線性時間(linear time)排序演算法的介紹。

Introduction

- How do we measure the goodness of an algorithm?
- How do we measure the difficulty of a problem?
- How do we know that an algorithm is optimal for a problem?
- How can we know that there does not exist any other better algorithm to solve the same problem?

Straight insertion sort

學習目標

- 插入排序 (insertion sort) 演算法的設計
- 插入排序 (insertion sort) 演算法的複雜度分析。

Straight insertion sort

input: 7, 5, 1, 4, 3

7, 5, 1, 4, 3

5, 7, 1, 4, 3

1, 5, 7, 4, 3

1, 4, 5, 7, 3

1, 3, 4, 5, 7

Algorithm Straight Insertion Sort

Input: $x_1, x_2, \ldots, x_n$

Output: The sorted sequence of $x_1, x_2, \ldots, x_n$

For $j := 2$ to n do

Begin

$i := j-1$

$x := x_j$

While $x < x_i$ and $i > 0$ do

Begin

$x_{i+1} := x_i$

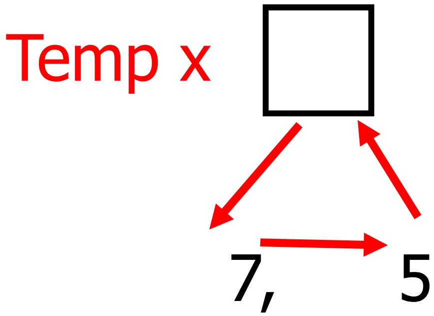
$i := i-1$

End

$x_{i+1} := x$

End

Always do



input: 7, 5, 1, 4, 3

7, 5, 1, 4, 3

5, 7, 1, 4, 3

1, 5, 7, 4, 3

1, 4, 5, 7, 3

1, 3, 4, 5, 7

Inversion table

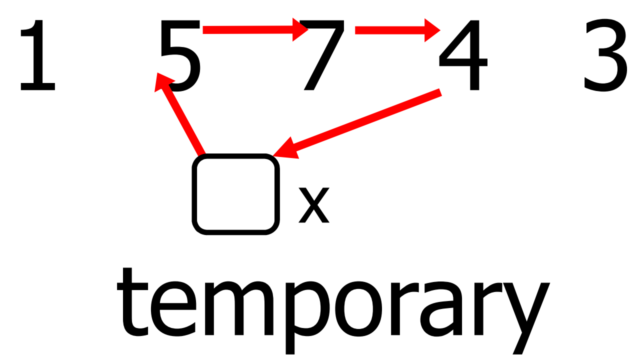
- $(a_1, a_2, \ldots, a_n)$: a permutation of $\{1, 2, \ldots, n\}$
- $(d_1, d_2, \ldots, d_n)$: the inversion table of $(a_1, a_2, \ldots, a_n)$
- d_i: the number of elements to the left of i that are greater than i
- e.g. permutation (7 5 1 4 3 2 6)

  inversion table 2 4 3 2 1 1 0
- e.g. permutation (7 6 5 4 3 2 1)

  inversion table 6 5 4 3 2 1 0
- d_i: the number of movements executed for x_i in the inner do loop.

Analysis of # of movements

- d_i: # of data movements in straight insertion sort



- e.g. $d_4=2$
- $$X = \sum_{i=2}^{n}(2+d_i) = 2(n-1) + \sum_{i=2}^{n}(d_i)$$

// for i = 2 to n

Analysis by inversion table

- best case: already sorted

  $d_i = 0$ for $1 \le i \le n$

  $\Rightarrow X = 2(n - 1) = O(n)$

A: (1, 2, 3, 4, 5, 6, 7)
D: (0, 0, 0, 0, 0, 0, 0)

- worst case: reversely sorted

  $d_1 = 0$

  $d_2 = 1$

  :

  $d_i = n - i$

  $d_n = n-1$

A: (7, 6, 5, 4, 3, 2, 1)
D: (6, 5, 4, 3, 2, 1, 0)

$$X = \sum_{i=2}^{n} (2 + d_i) = 2(n-1) + \frac{n(n-1)}{2} = O(n^2)$$

- <u>**average case**</u>**:**

  x_i is being inserted into the sorted sequence

  $x_1\ x_2\ \ldots\ x_{i-1}$

- the probability that x_i is the largest: 1/i
  - takes 2 data movements ($2+d_i=2$, $d_i=0$)
- the probability that x_i is the second largest : 1/i
  - takes 3 data movements

1 4 7 5

- # of movements for inserting x_j:

$$2 + d_i = \frac{2}{i} + \frac{3}{i} + \qquad + \frac{i+1}{i} = \sum_{j=1}^{i} \frac{j+1}{i} = \frac{i+3}{2}$$

$$X = \sum_{i=2}^{n} \frac{i+3}{2} = \frac{1}{2}\left(\sum_{i=2}^{n} i + \sum_{i=2}^{n} 3 \right) = \frac{(n+8)(n-1)}{4} = O(n^2)$$

Summation Formula

$$\sum_{i=0}^{n} i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

$$\sum_{i=0}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$$

$$\sum_{i=0}^{n} i^3 = \left[\sum_{i=0}^{n} i\right]^2 = \left[\frac{n(n+1)}{2}\right]^2 = \frac{n^4}{4} + \frac{n^3}{2} + \frac{n^2}{4}$$

$$\sum_{i=0}^{n} i^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30} = \frac{n^5}{5} + \frac{n^4}{2} + \frac{n^3}{3} - \frac{n}{30}$$

$$\sum_{i=1}^{n} i^5 = \frac{n^2(n+1)^2(2n^2+2n-1)}{12}$$

$$\sum_{i=1}^{n} i^6 = \frac{n(n+1)(2n+1)(3n^4+6n^3-3n+1)}{42}$$

$$\sum_{i=1}^{n} i^7 = \frac{n^2(n+1)^2(3n^4+6n^3-n^2-4n+2)}{24}$$

$$\sum_{i=1}^{n} \log i = \log n! \text{ (The property of logarithms)}$$

$$\sum_{i=2}^{n} 2\ln\left(\frac{i}{i-1}\right) = 2\ln(n)$$

$$\sum_{i=2}^{n} i\ln\left(\frac{i}{i-1}\right) + \ln(i-1) = n\ln(n)$$

$$\sum_{i=1}^{n} 2\log\left(\frac{1}{i}\right) = 2\log\left(\frac{1}{n!}\right)$$

$$\sum_{i=0}^{n-1} a^i = \frac{1-a^n}{1-a}$$

$$\sum_{i=0}^{n-1} \frac{1}{2^i} = 2 - \frac{1}{2^{n-1}}$$

$$\sum_{i=0}^{n-1} ia^i = \frac{a - na^n + (n-1)a^{n+1}}{(1-a)^2}$$

$$\sum_{i=0}^{n-1} i2^i = 2 + (n-2)2^n$$

$$\sum_{i=0}^{n-1} \frac{i}{2^i} = 2 - \frac{n+1}{2^{n-1}}$$

Summation formula

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n$$

$$\sum_{k=0}^{m} \binom{n+k}{n} = \binom{n+m+1}{n+1}$$

$$\sum_{i=1}^{n} i\binom{n}{i} = n(2^{n-1})$$

$$\sum_{i=0}^{n} {}_iP_k \binom{n}{i} = {}_nP_k(2^{n-k})$$

$$\sum_{i=0}^{n} \frac{\binom{n}{i}}{i+1} = \frac{2^{n+1}-1}{n+1}$$

Summation formula

$$\sum_{i=0}^{n} i! \cdot \binom{n}{i} = \sum_{i=0}^{n} {}_nP_i = \lfloor n! \cdot e \rfloor, \quad n \in \mathbb{Z}^+,$$

$$\sum_{i=k}^{n} \binom{i}{k} = \binom{n+1}{k+1}$$

$$\sum_{i=0}^{n} \binom{n}{i} a^{n-i} b^i = (a+b)^n,$$

$$\sum_{i=0}^{n} i \cdot i! = (n+1)! - 1$$

$$\sum_{i=1}^{n} {}_{i+k}P_{k+1} = \sum_{i=1}^{n} \prod_{j=0}^{k} (i+j) = \frac{(n+k+1)!}{(n-1)!(k+2)}$$

$$\sum_{i=0}^{n} \binom{m+i-1}{i} = \binom{m+n}{n}$$

Analysis of # of exchanges

- <u>Method 1 (straightforward)</u>

- x_i is being inserted into the sorted sequence

  $x_1\ x_2\ \ldots.\ x_{i-1}$

- If x_j is the kth ($1 \le k \le i$) largest, it takes (k–1) exchanges.

- e.g. 1 5 7↔4

  1 5↔4 7

  1 4 5 7

- # of exchanges required for x_i to be inserted:

$$\frac{0}{i} + \frac{1}{i} + \qquad + \frac{i-1}{i} = \frac{i-1}{2}$$

- # of exchanges for sorting:

$$\sum_{i=2}^{n} \frac{i-1}{2}$$

$$= \sum_{i=2}^{n} \frac{i}{2} - \sum_{i=2}^{n} \frac{1}{2}$$

$$= \frac{1}{2} \cdot \frac{(n-1)(n+2)}{2} - \frac{n-1}{2}$$

$$= \frac{n(n-1)}{4}$$

Question:

- Which is the inversion table of the array of the input (7, 5, 1, 4, 3, 2, 6)?

(1) (6, 5, 4, 3, 2, 1, 0)

(2) (0, 0, 0, 0, 0, 0, 0)

(3) (0, 1, 2, 3, 4, 5, 6)

(4) (2, 4, 3, 2, 1, 1, 0).

Ans. 4

Binary search

學習目標

- Binary search演算法的設計
- Binary search演算法的複雜度分析。

Binary search

- sorted sequence : (search 9)

	1	4	5	7	9	10	12	15
step 1				↑				
step 2						↑		
step 3					↑			

- best case: 1 step = $O(1)$
- worst case: $(\lfloor \log_2 n \rfloor + 1)$ steps = $O(\log n)$
- average case: $O(\log n)$ steps

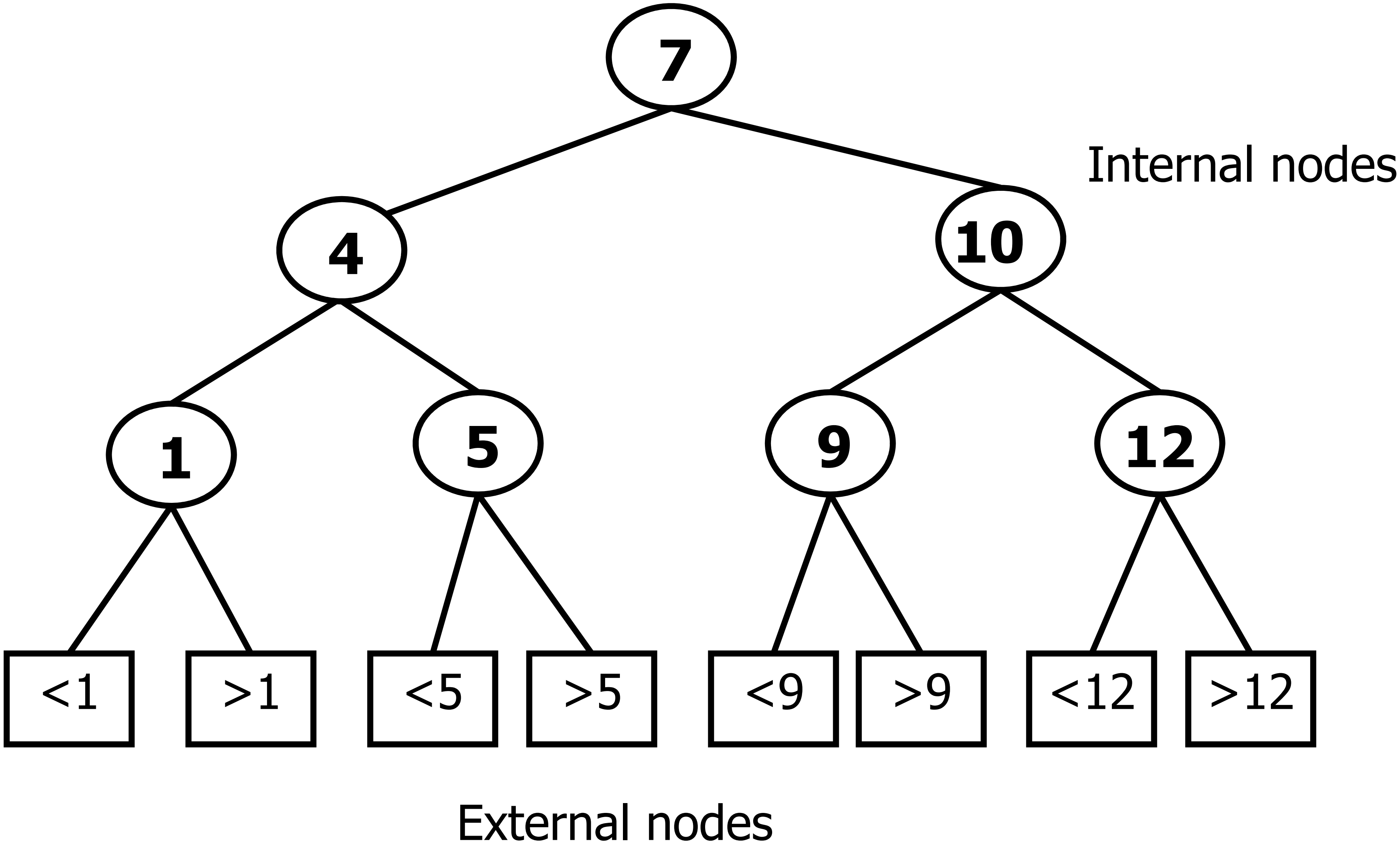
Binary Search Algorithm

Input : $a_1, a_2, \ldots, a_n$, $n > 0$, with $a_1 \le a_2 \le \ldots \le a_n$, and x

Output : j if $a_j = X$ and 0 if no j exists such that $a_j = X$.

```
i := 1
m := n
while (i ≤ m) do
      begin  j := ⌊(i+m)/2⌋
                if (x = a_j) then output j & stop
                if (x < a_j)  then m := j–1
                else i := j+1
      end
j := 0
output j
```

Binary Searching Tree



The binary Search (Analysis-Average case) ＊

找得到的情況：

計有 1 個情況，是找了	1 次即得
2	2
4	3
：	：
2^{i-1}	i

The binary Search (Analysis-Average case)

＊ 找不到的情況：

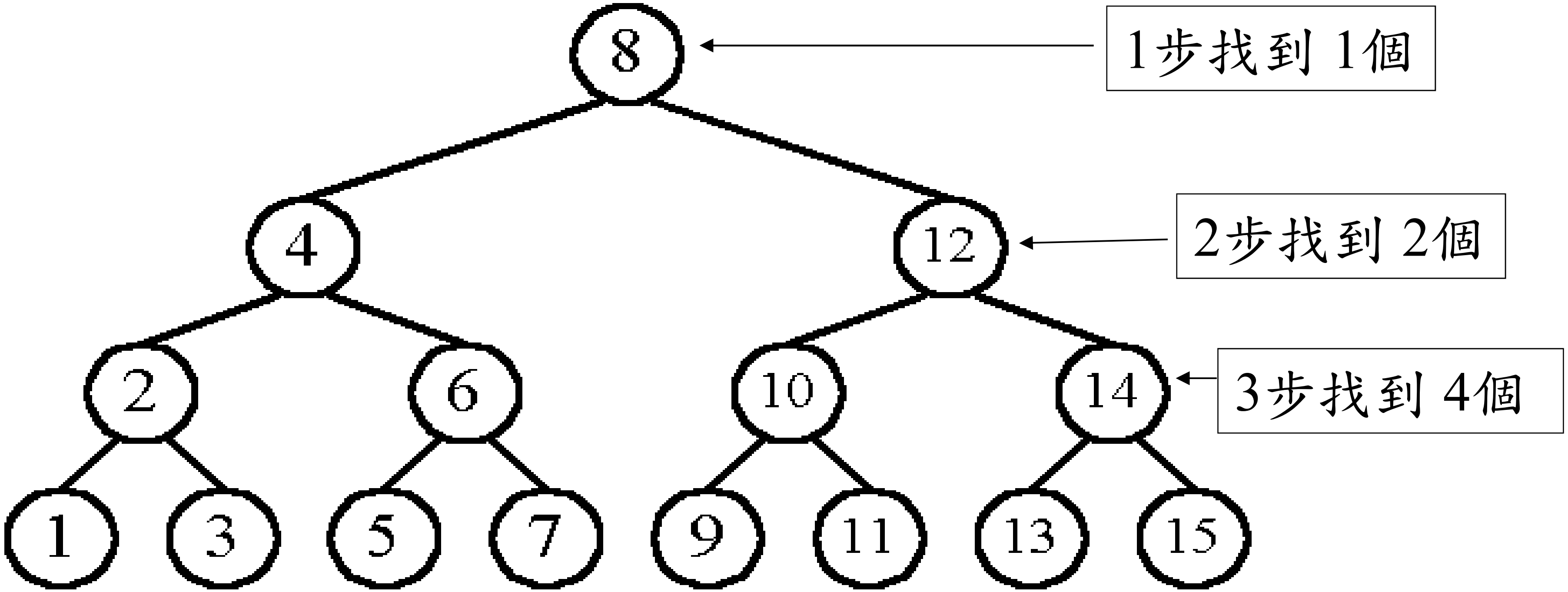
- 在 $(n+1)$ 種情況裡(外部節點的個數)，每一種都得找 $\lfloor \log n \rfloor + 1$ 次方可確定。

$\therefore$ 平均 "找" 的次數 $A(n) = \dfrac{1}{2n+1}\left(\sum_{i=1}^{k} i \cdot 2^{i-1} + k(n+1)\right)$

（令 $k = \lfloor \log n \rfloor + 1$）

利用歸納法 (induction) 可得：

$A(n) < k = O(\lfloor \log n \rfloor)$



i步找到2^{i-1}個

n cases for successful search

n+1 cases for unsuccessful search

Assume n=2^k-1個

Average # of comparisons done in the binary tree:

$$A(n) = \frac{1}{2n+1}\left(\sum_{i=1}^{k} i\,2^{i-1} + k(n+1)\right), \text{ where } k = \lfloor \log n \rfloor + 1$$

K步找不到 (n+1)個

Assume $n=2^k$

$$\sum_{i=1}^{k} i\,2^{i-1} = 2^k(k-1)+1 \qquad (2\text{-}1)$$

proved by induction on k (skip, ref. p.25)

Assume $n=2^k-1$個, $n+1=2^k$

$$A(n) = \frac{1}{2n+1}\left(\sum_{i=1}^{k} i\,2^{i-1} + k(n+1)\right)$$

$$A(n) = \frac{1}{2n+1}((k-1)2^k+1+k(2^k))$$

$$A(n) \approx \frac{1}{2^{k+1}}(2^k(k-1)+1+k2^k)$$

$$= \frac{(k-1)}{2} + \frac{k}{2} = k - \frac{1}{2}$$

$\approx k = \log n = O(\log n)$ as n is very large

Question:

- Which problem the worst-case time complexity of the binary search on a sorted sequences with n elements?

(1) O(1)

(2) O(log n)

(3) O(n)

(4) O(nlog n).

Ans. 2

Straight selection sort

學習目標

- 選擇排序 (selection sort) 演算法的設計
- 選擇排序 (selection sort) 演算法的複雜度分析。

Straight selection sort

- Find the smallest number.
- Let this smallest number occupy a_1 by exchanging a_1 with this smallest number.
- Repeat the above step on the remaining numbers. That is, find the second smallest number and let it occupy a_2.
- Continue the process until the largest number is found.

Straight Selection Sort

- **Input:** $a_1, a_2, \ldots, a_n$.
- **Output: The sorted sequence of** $a_1, a_2, \ldots, a_n$.

For j **:=1 to** $n-1$ **do**

Begin

$\boxed{f := j}$

Flag used to point the Smallest element

For k **:=** j**+1 to n do**

If $a_k < a_f$ **then** $f := k$

$a_j \leftrightarrow a_f$

End

Two operations:
(1) comparison
(2) change flag

Straight selection sort

7 5 1 4 3
1 5 7 4 3
1 3 7 4 5
1 3 4 7 5
1 3 4 5 7

First run
7>5 change
5>1 change
1<4 no change
1<3 no change

- **The number of comparisons** of two elements is a fixed number; namely n(n-1)/2. That is, no matter what the input data are, we always have to perform *n(n-1)/2* comparisons.
- Only consider # of changes in the flag which is used for selecting the smallest number in each iteration.
  - best case: O(1) sorted sequence
  - worst case: $O(n^2)$
  - average case: O(n log n)

of changing flags

- Define $f(a_1, a_2, \ldots., a_n)$ denote the number of changing flags need to find the smallest number of the permutation $f(a_1, a_2, \ldots, a_n)$. ***找出最小數所需改變 flag 次數***
- n=2

  (1, 2) 0 次 or (2, 1) 1 次
- n=3

a_1	a_2	a_3	$f(a_1, a_2, a_3)$
1	2	3	0
1	3	2	0
2	1	3	1
2	3	1	1
3	1	2	1
3	2	1	2

Determine $f(a_1, a_2, \ldots, a_n)$

- If $a_n=1$, then $f(a_1, a_2, \ldots, a_n)=1+f(a_1, a_2, \ldots, a_{n-1})$ because it must be a change of flag at the last flag.
- If $a_n \neq 1$, then $f(a_1, a_2, \ldots, a_n)=f(a_1, a_2, \ldots, a_{n-1})$ because it must be no change of flag at the last flag.
- **Let $P_n(k)$ denote that the probability of a permutation $\{a_1, a_2, \ldots, a_n\}$ need *k* changes to find the smallest number**.
- For example:
  - $P_3(0)=2/6$,
  - $P_3(1)=3/6$,
  - $P_3(2)=1/6$

- The average number of changing flags to find the smallest number is

$$X_n = \sum_{k=0}^{n-1} k P_n(k)$$

- The average number of changing flags

$$A(n) = X_n + A(n-1)$$

- $P_n(k)=P(a_n=1)\ P_{n-1}(k-1)+P(a_n\neq 1)\ P_{n-1}(k)$

Solve $P_n(k)=P(a_n=1)\ P_{n-1}(k-1)+P(a_n\neq 1)\ P_{n-1}(k)$

- *Since* $P(a_n=1) = 1/n$ *and* $P(a_n\neq 1)=(n-1)/n$
- *Therefor*

$$P_n(k)=1/n\ P_{n-1}(k-1)+(n-1)/n\ P_{n-1}(k)$$

- *Furthermore*

$$P_n(k) = \begin{cases} 1, & if\ k = 0 \\ 0, & if\ k \neq 0 \\ 0, & if\ k < 0\ and\ if\ \mathrm{k} = n \end{cases}$$

For example,

$P_2(0)=1/2,$

$P_2(1)==1/2$

$P_n(k)=1/n\ P_{n-1}(k-1)+(n-1)/n\ P_{n-1}(k)$

And

$P_3(0)= 1/3\ P_2(-1) + 2/3\ P_2(0) =1/3 \times 0 + 2/3 \times ½ =1/3$

$P_3(2)= 1/3\ P_2(1) + 2/3\ P_2(2) =1/3 \times 1/2 + 2/3 \times 0 =1/6$

$$\boxed{X_n = \sum_{k=1}^{n-1} k P_n(k) = \frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{2} = H_n - 1, \qquad (2.4)}$$

Proof by Induction (skip)

Solve $A(n)=X_n+A(n-1)$

- $A(n)$

$=X_n+A(n-1)$

$= (H_n-1)+A(n-1)$

$= (H_n-1)+ (H_{n-1}-1)+ (H_{n-2}-1)+\ldots + (H_2-1)$

$$= \sum_{i=2}^{n} H_i - (n-1)$$

$A(n) \quad = X_n+A(n-1)$
$A(n-1)= X_{n-1}+A(n-2)$
$A(n-2)= X_{n-2}+A(n-3)$
$A(n-3)= X_{n-3}+A(n-4)$

$A(2) = X_2+A(1)$

$A(n)=X_n+X_{n-1}+\ldots+X_2+A(1)$
$=(H_n-1)+(H_{n-1}-1)+(H_{n-2}-1)+ \ldots+(H_2-1)$

$$\sum_{i=1}^{n} H_i = H_n + H_{n-1} + \ldots + H_1$$

$$= H_n + (H_n - \frac{1}{n}) + \ldots + (H_n - \frac{1}{n} - \frac{1}{n-1} - \ldots - \frac{1}{2})$$

$$= nH_n - (\frac{n-1}{n} + \frac{n-2}{n-1} + \ldots + \frac{1}{2})$$

$$= nH_n - (1 - \frac{1}{n} + 1 - \frac{1}{n-1} + \ldots + 1 - \frac{1}{2})$$

$$= nH_n - (n - 1 - \frac{1}{n} - \frac{1}{n-1} - \ldots - \frac{1}{2})$$

$$= nH_n - n + H_n$$

$$= (n+1)H_n - n \, .$$

Straight selection sort

$$\sum_{i=1}^{n} H_i = (n+1)H_i - n \qquad \sum_{i=2}^{n} H_i = (n+1)H_i - H_1 - n$$

$$\mathrm{A(n)} = \sum_{i=2}^{n} H_i - (n-1)$$

$= (n+1)H_i - H_1 - n - (\mathrm{n} - 1)$

$= (n+1)H_n - 2n$

$$\boxed{\begin{array}{l} 1 + \frac{n}{2} \le H_{2^n} \le 1 + n \\ H_k \le 1 + \log_2 K \end{array}}$$

Therefore:

A(n)=O(n log n)

Question:

- Which is the value of *f(3,2,1)* in the straight selection sort?

(1) 0

(2) 1

(3) 2

(4) 3.

Ans. 3

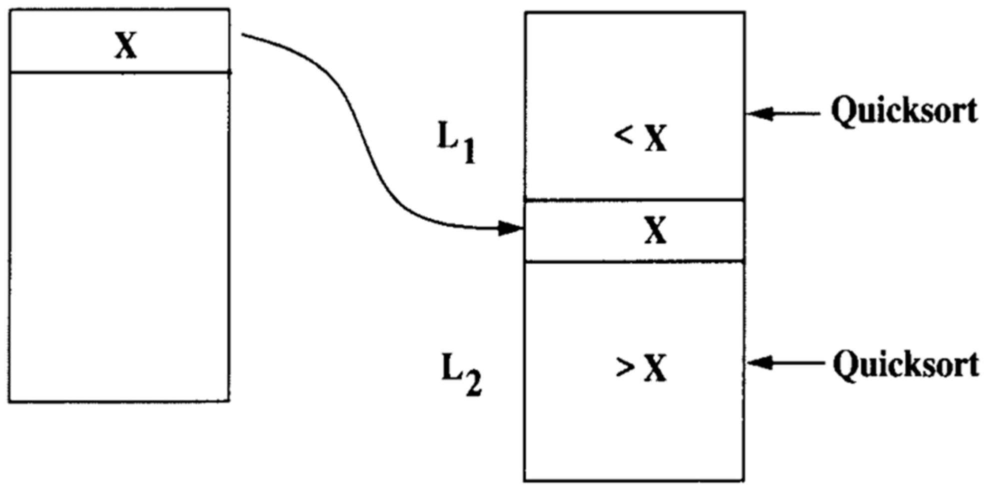
QuickSort

學習目標

- 快速排序 (quicksort) 演算法的設計
- 快速排序 (quicksort) 演算法的複雜度分析。

QuickSort

- Quicksort is based upon the divide-and-conquer strategy.
- Divide-and-conquer strategy divides a problem into two sub-problems and solves these two subproblems individually and independently. We later merge the results.
- Given a set of numbers $a_1, a_2, \ldots, a_n$ we choose an element X to divide $a_1, a_2, \ldots, a_n$ into two lists.
- After the dividing, we may apply Quicksort to both L_1 and L_2 recursively and the resulting list is a sorted list.



Quicksort

11	5	24	2	31	7	8	26	10	15
11	5	10	2	31	7	8	26	24	15
11	5	10	2	8	7	31	26	24	15
7	5	10	2	8	11	31	26	24	15

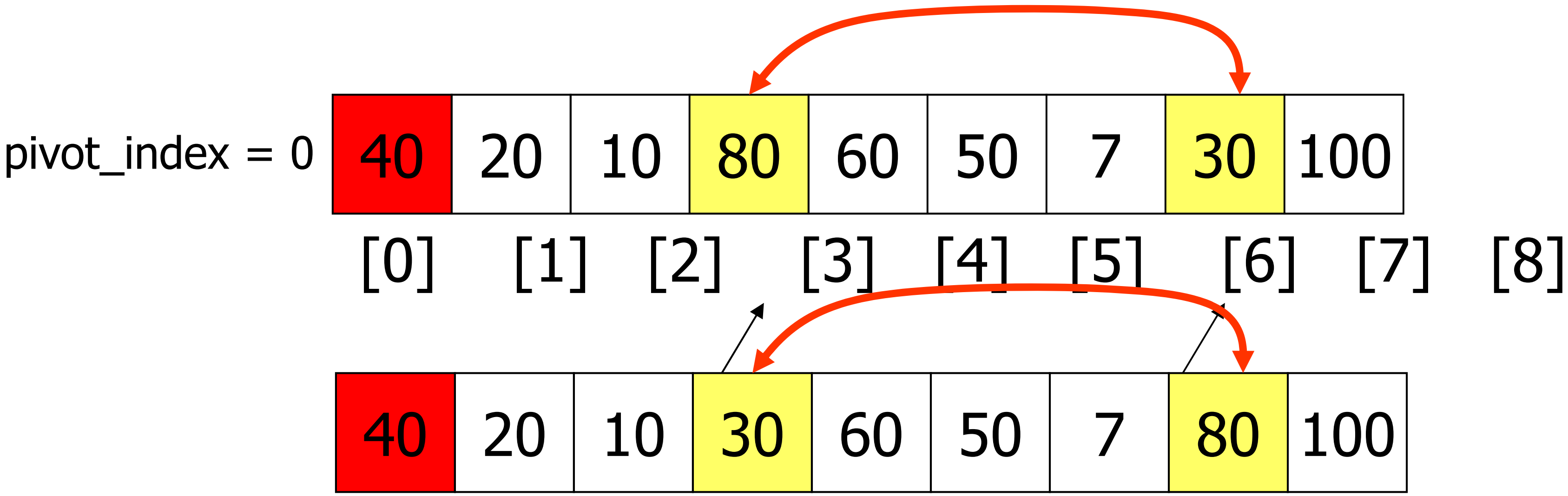
$|\leftarrow \quad <11 \quad \rightarrow|$ $\qquad$ $|\leftarrow \quad >11 \quad \rightarrow|$

- Recursively apply the same procedure.

Quicksort

40	20	10	80	60	50	7	30	100

- Given a *pivot*, partition the elements of the array such that the resulting array consists of:
  - One sub-array that contains elements ≥ pivot (if not distinct numbers)
  - Another sub-array that contains elements < pivot



40	20	10	30	60	50	7	80	100

40	20	10	30	7	50	60	80	100

pivot_index = 0

40	20	10	30	7	50	60	80	100
[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]

too_big_index too_small_index

pivot_index = 4

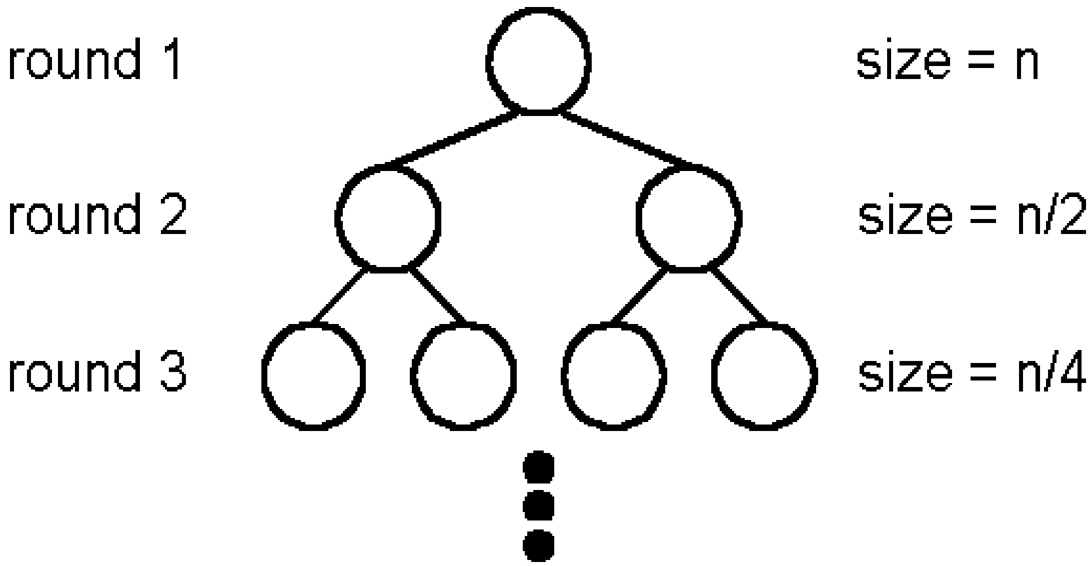
7	20	10	30	40	50	60	80	100
[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]

too_big_index too_small_index

7	20	10	30	40	50	60	80	100

Best case of Quicksort

- Best case: *O(nlogn)*
- A list is split into two sublists with almost equal size.



- *log n* rounds are needed
- In each round, *n* comparisons (ignoring the element used to split) are required.

Worst case of Quicksort

- Worst case: $O(n^2)$
- Sorted or reverse sorted.
- In each round, the number used to split is either the smallest or the largest.

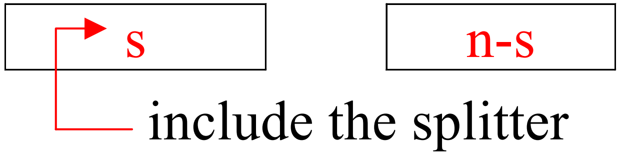
(1, 2, 3, 4, 5, 6, 7)

$$n + (n-1) + (n-2) + \cdots + 1$$
$$= \frac{n(n-1)}{2}$$
$$= O(n^2)$$

(7, 6, 5, 4, 3, 2, 1)

Average case of Quicksort

- Average case: O(n log n)



The number of operations needed for first splitting operation

$$
\begin{aligned}
T(n) &= \mathop{\mathrm{Avg}}_{1\le s\le n}\left(T(s)+T(n-s)\right)+\boxed{cn} \\
&= \frac{1}{n}\sum_{s=1}^{n}\left(T(s)+T(n-s)\right)+cn \\
&= \frac{1}{n}\left(T(1)+T(n-1)+T(2)+T(n-2)+\cdots+T(n)+T(0)\right)+cn,\ T(0)=0 \\
&= \frac{1}{n}\left(2T(1)+2T(2)+\cdots+2T(n-1)+T(n)\right)+cn
\end{aligned}
$$

$$(n-1)T(n) \quad = 2T(1)+2T(2)+\cdots+2T(n-1) + cn^2\cdots\cdots(1)$$

$$(n-2)T(n\text{-}1) = 2T(1)+2T(2)+\cdots+2T(n-2)+c(n-1)^2\cdots(2)$$

Let n=n-1 to (1)

$$(1) - (2)$$

$$(n-1)T(n) -(n-2)T(n-1) = 2T(n-1)+c(2n-1)$$

$$(n-1)T(n) -nT(n-1) = c(2n-1)$$

Divide by n(n-1)

C(2n-1)/n(n-1)= A/n + B/(n-1)
Find A and B

部份分式

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

$$= c(\frac{1}{n} + \frac{1}{n-1}) + c(\frac{1}{n-1} + \frac{1}{n-2}) + \cdots + c(\frac{1}{2} + 1) + T(1),\ T(1) = 0$$

$$= c(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{2}) + c(\frac{1}{n-1} + \frac{1}{n-2} + \ldots + 1)$$

$$= c(H_n - 1) + cH_{n-1}$$

Harmonic number [Knuth 1986]

$$H_n = 1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}$$

$$=\ln n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} - \varepsilon, \text{ where } 0<\varepsilon<\frac{1}{252n^6}$$

$$\gamma = 0.5772156649\cdots.$$

$$H_n = O(\log n)$$

$$T(\mathrm{n})/n = \mathrm{c}(H_n\text{-}1) + cH_{n-1}$$

$$\Rightarrow T(\mathrm{n})/\mathrm{n} = \mathrm{c}(H_n + H_n\text{-}1\text{-}\ 1/n)$$

$$\Rightarrow T(n) = 2\mathrm{c}nH_n - \mathrm{c}(\mathrm{n}+1)$$

$$\Rightarrow \qquad = O(n \log n)$$

Question:

- For the input array (40, 20, 10, 80, 60, 50, 7, 30, 100) stored in array a[0..8], what is the index of first element 40 after performing quicksort?

(1) 0

(2) 1

(3) 2

(4) 3.

Ans. 4

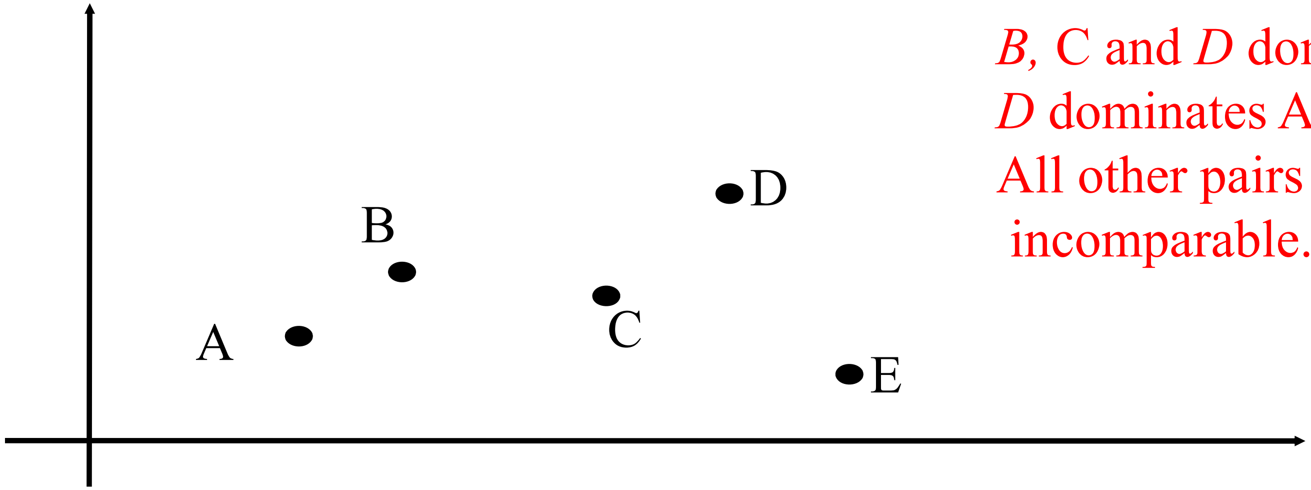
2-D ranking finding

學習目標

- 2-D ranking finding problem 問題定義
- 2-D ranking finding problem演算法的設計
- 2-D ranking finding problem 演算法的複雜度分析。

2-D ranking finding

- **Def**: Let $A = (a_1, a_2)$, $B = (b_1, b_2)$. A dominates B iff $a_1 > b_1$ and $a_2 > b_2$
- **Def**: If neither A dominates B nor B dominates A, then A and B are incomparable.
- **Def**: Given a set S of n points, the rank of a point x is the number of points dominated by x.



B, C and *D* dominate *A.*
D dominates A, B and C.
All other pairs of points are incomparable.

rank(A)= 0 rank(B) = 1 rank(C) = 1 rank(D) = 3 rank(E) = 0

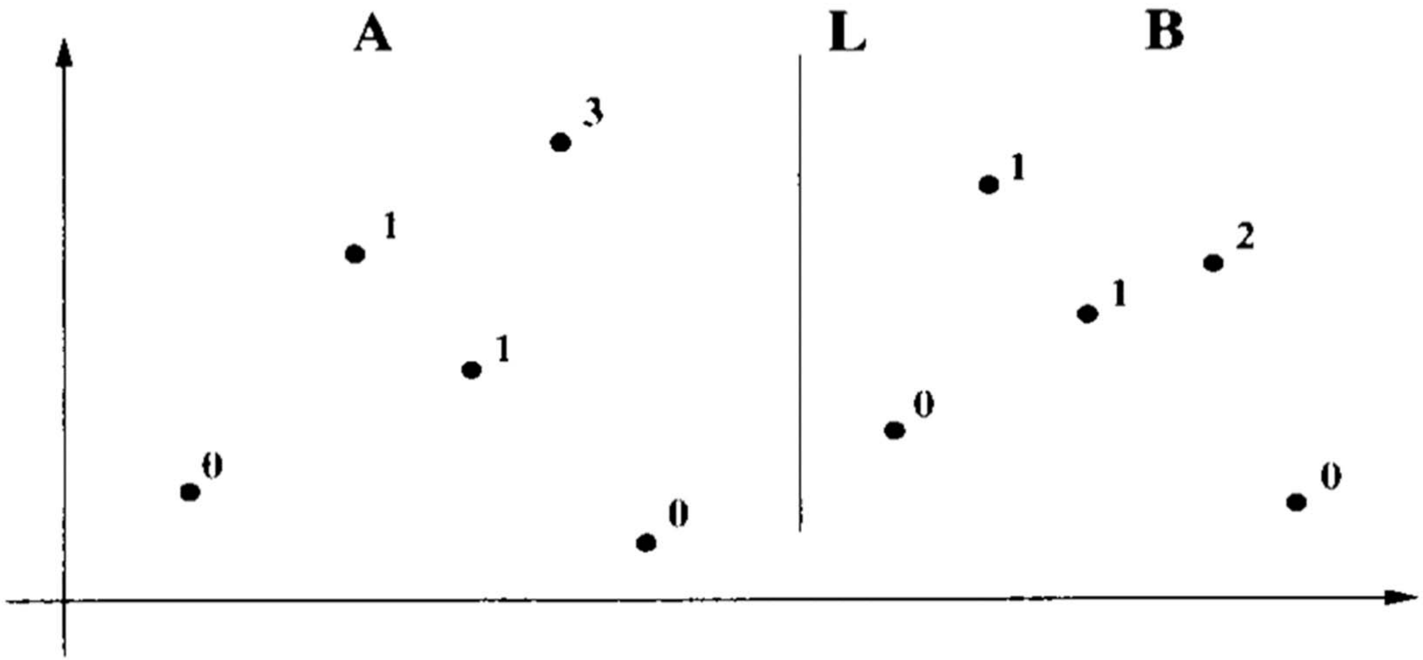
Rank Finding Problem

- **Find the rank of every points.**
- Straightforward algorithm**:**
  - compare all pairs of points : $O(n^2)$

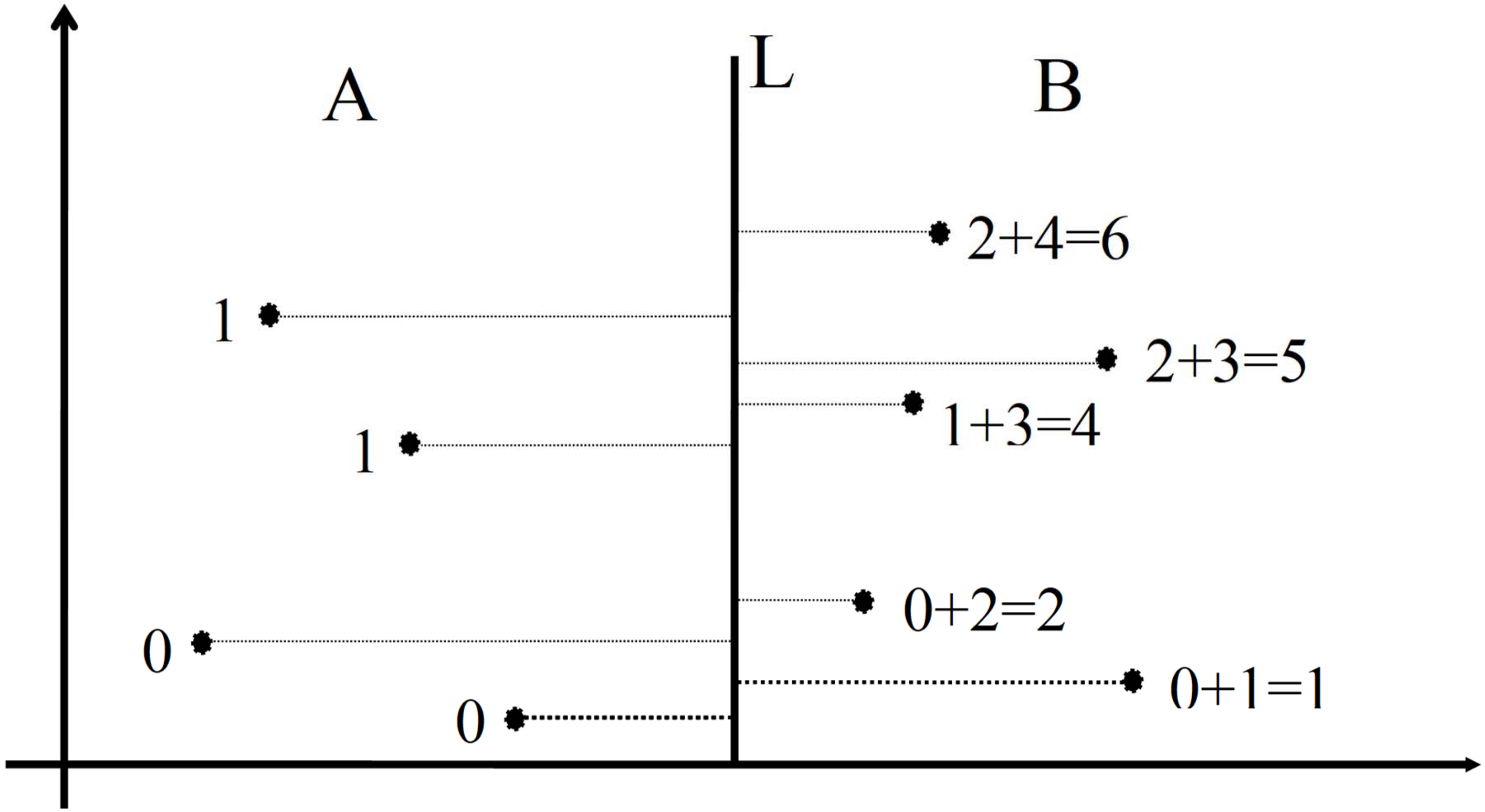
- **Divide-and-conquer** 2-D ranking finding
  - Step 1: Split the points along the median line L into A and B.
  - Step 2: Find ranks of points in A and ranks of points in B, recursively.
  - Step 3: Sort points in A and B according to their y-values. Update the ranks of points in B.

Local ranks before merge

- Find a straight line L perpendicular to the x-axis which separates the set of points into two subsets and these two subsets are of equal size.
- The rank of any point in A **will not** be affected by the presence of B.
- But the rank of a point in B may be affected the presence of A.



- divide-and-conquer algorithm



Algorithm 2–5 □ A rank finding algorithm

Input: A set S of planar points $P_1, P_2, \ldots, P_n$.

Output: The rank of every point in S.

Step 1. If S contains only one point, return its rank as 0. Otherwise, choose a cut line L perpendicular to the x-axis such that $n/2$ points of S have X-values less than L (call this set of points A) and the remainder points have X-values greater than L (call this set B). Note that L is a median X-value of this set.

Step 2. Recursively, use this rank finding algorithm to find the ranks of points in A and ranks of points in B.

Step 3. Sort points in A and B according to their y-values. Scan these points sequentially and determine, for each point in B, the number of points in A whose y-values are less than its y-value. The rank of this point is equal to the rank of this point among points in B (found in Step 2), plus the number of points in A whose y-values are less than its y-value.

- time complexity : step 1 : $O(n)$ (finding median)
  step 3 : $O(n \log n)$ (sorting)

- total time complexity :

$$
\begin{aligned}
T(n) &\le 2T(\tfrac{n}{2}) + c_1 n \log n + c_2 n \\
&\le 2T(\tfrac{n}{2}) + c\, n \log n \\
&\le 4T(\tfrac{n}{4}) + c\, n \log \tfrac{n}{2} + c\, n \log n \\
&\le nT(1) + c(n \log n + n \log \tfrac{n}{2} + n \log \tfrac{n}{4} + \cdots + n \log 2) \\
&= nT(1) + \frac{cn \log n(\log n + \log 2)}{2} \\
&= O(n \log^2 n)
\end{aligned}
$$

If $n = 2^p$, $p = \log n$

$$
\begin{aligned}
&\log n + \log \frac{n}{2} + \log \frac{n}{4} + \ldots + \log 2 \\
&= p + (p-1) + (p-2) + \ldots + 1 \\
&= \frac{p(p+1)}{2} = \log n(\log n + 1) = (\log n)^2
\end{aligned}
$$

For average & worst case

Question:

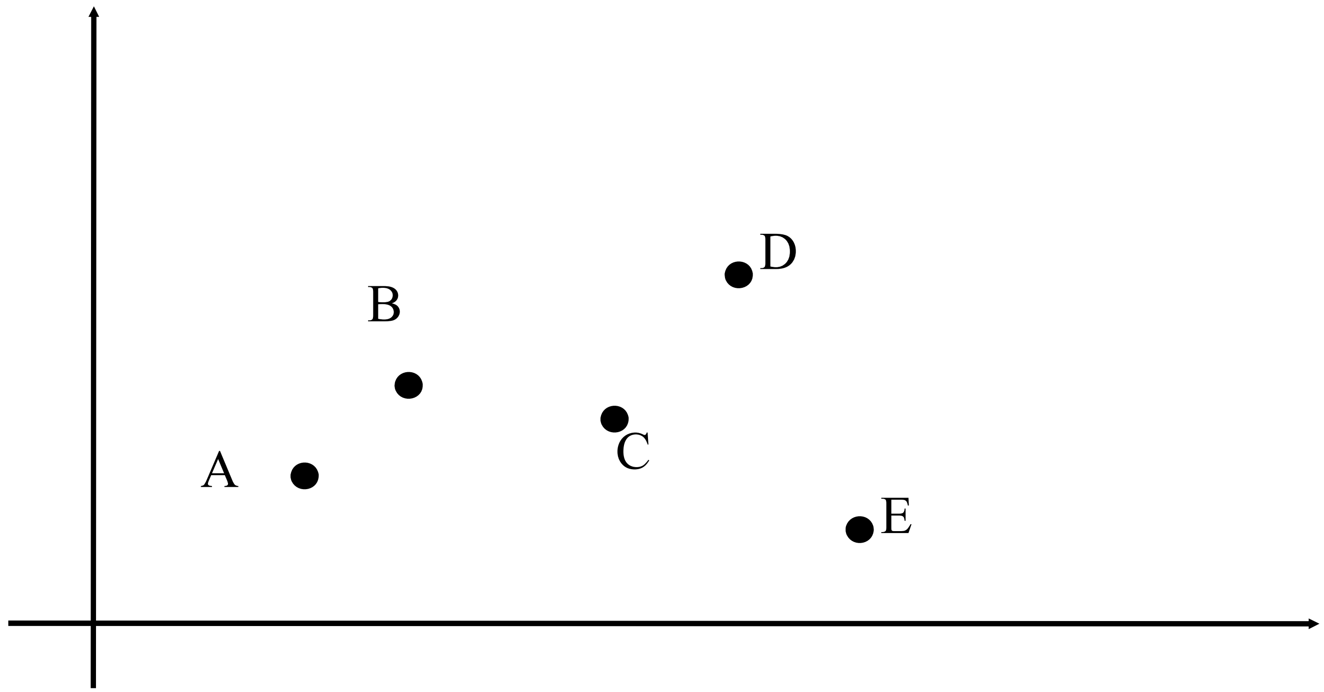
- For the input shown in follows, what is the value of rank(D)?

(1) 0

(2) 1

(3) 2

(4) 3.



Ans. 4

Lower bound

學習目標

- Lower Bound (LB) 定義
- Decision Tree 定義
- 排序演算法worst case lower bound 的複雜度分析。

Lower bound

- **How to we measure the difficulty of a problem?**
- **Def** : A lower bound of a problem is the least time complexity required for any algorithm which can be used to solve this problem.
- ☆ **worst case lower bound**

  ☆ **average case lower bound**
- Def : $f(n) = \Omega(g(n))$ "at least", "lower bound"

  $\exists$ c, and n_0, $\ni |f(n)| \geq c|g(n)| \ \forall \ n \geq n_0$

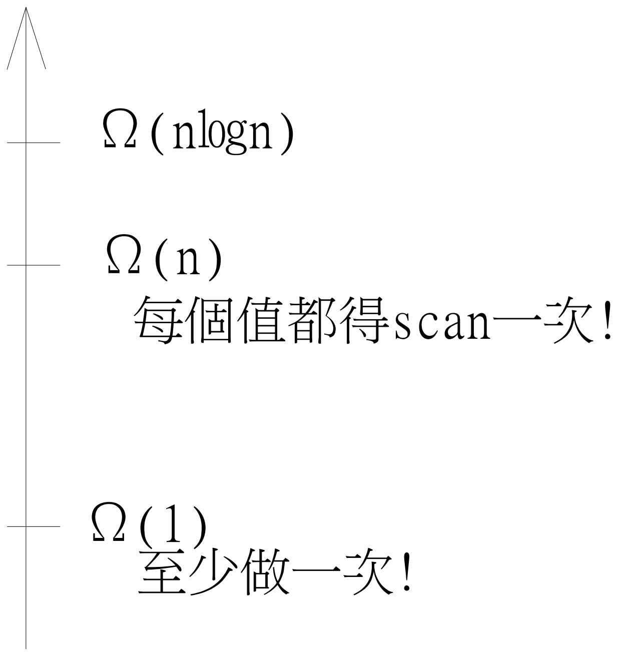
  e. g. $f(n) = 3n^2 + 2 = \Omega(n^2)$ or $\Omega(n)$
- The lower bound for a problem is not unique.
  - e.g. $\Omega(1)$, $\Omega(n)$, $\Omega(n \log n)$ are all lower bounds for sorting.
  - ($\Omega(1)$, $\Omega(n)$ are trivial)

Trivial lower bound

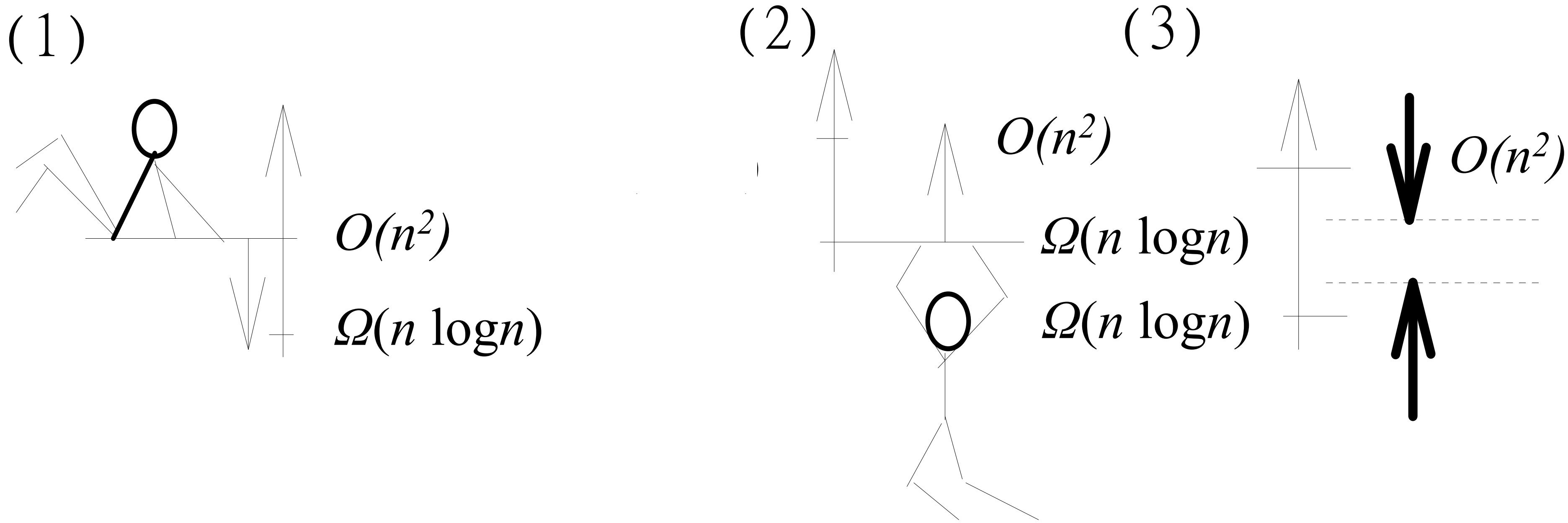
- ex.: sorting

$\Omega(1)$, $\Omega(n)$均為 trivial lower bound，討論它們沒有意義！

$\Omega(n^2)$如何？已有heapsort 其worst case為$\Omega(n\log n)$由 Def. 可知 lower bound 必須是所有 algorithms 中最小者，所以 $\Omega(n^2)$ 也不對！lower bound 至多是$\Omega(n\log n)$。



- 若目前 problem 之 highest lower bound 為 $\Omega(n \log n)$ 而找到的algorithm 最快的是$O(n^2)$，則：



- 若問題的 lower bound 為$\Omega(n \log n)$且找到的 algorithm 的 time-complexity 為 $O(n \log n)$

則 optimal algorithm of this problem 即已找到！

lower bound 與 algorithm 都無法再 improve。

The worst case lower bound of sorting

The worst case lower bound of sorting

- **Execution of an algorithm can be represented as binary trees.**
- **In general, any sorting algorithm whose basic operation is <u>compare</u> and <u>exchange</u> operation can be described by a binary tree.**
- **Straight insertion sort.**

6 permutations for 3 data elements

a_1	a_2	a_3
1	2	3
1	3	2
2	1	3
2	3	1
3	1	2
3	2	1

Straight insertion sort

- input data: (2, 3, 1)

  (1) a_1:a_2

  (2) a_2:a_3, $a_2 \leftrightarrow a_3$

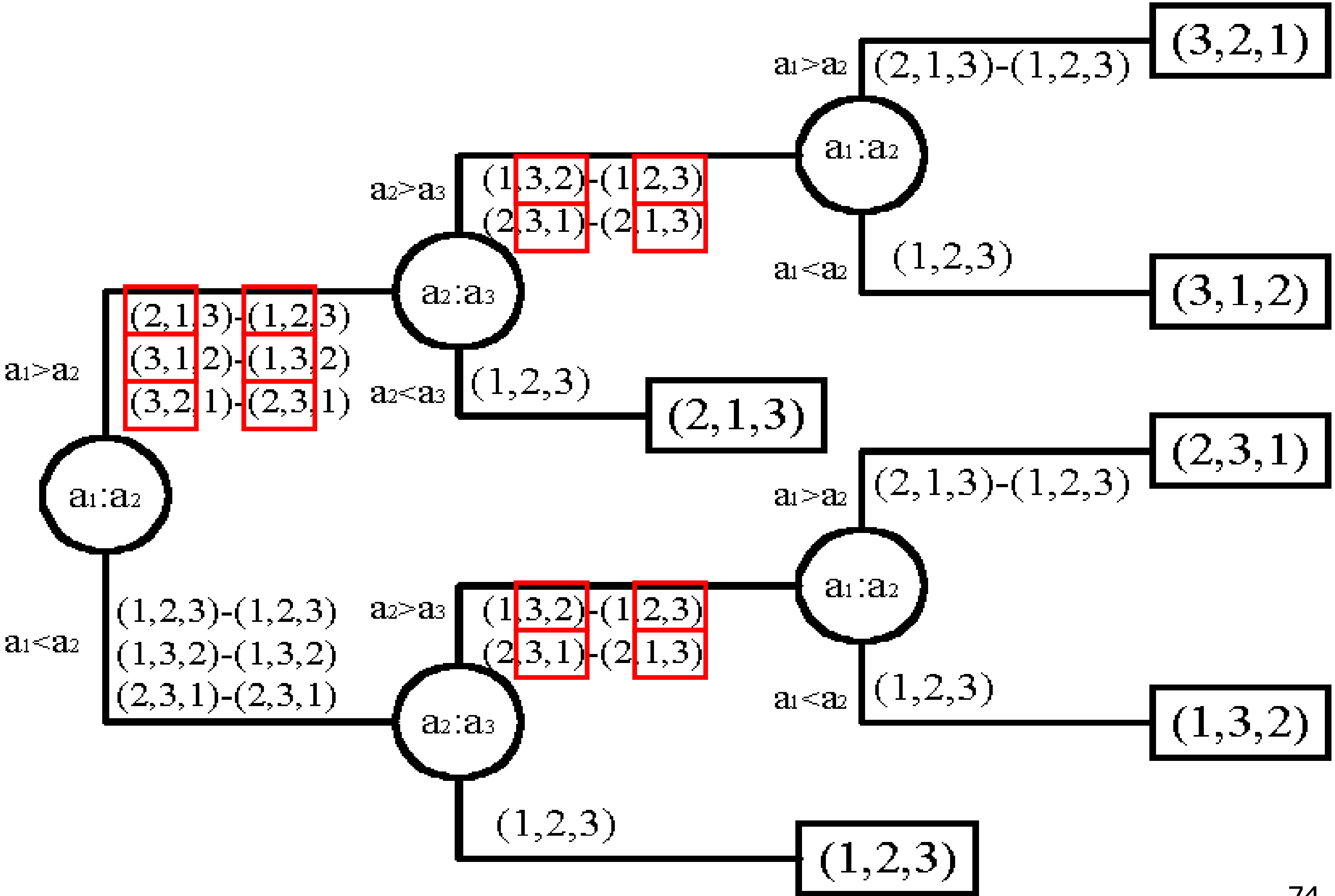
  (3) a_1:a_2, $a_1 \leftrightarrow a_2$

- input data: (2, 1, 3)

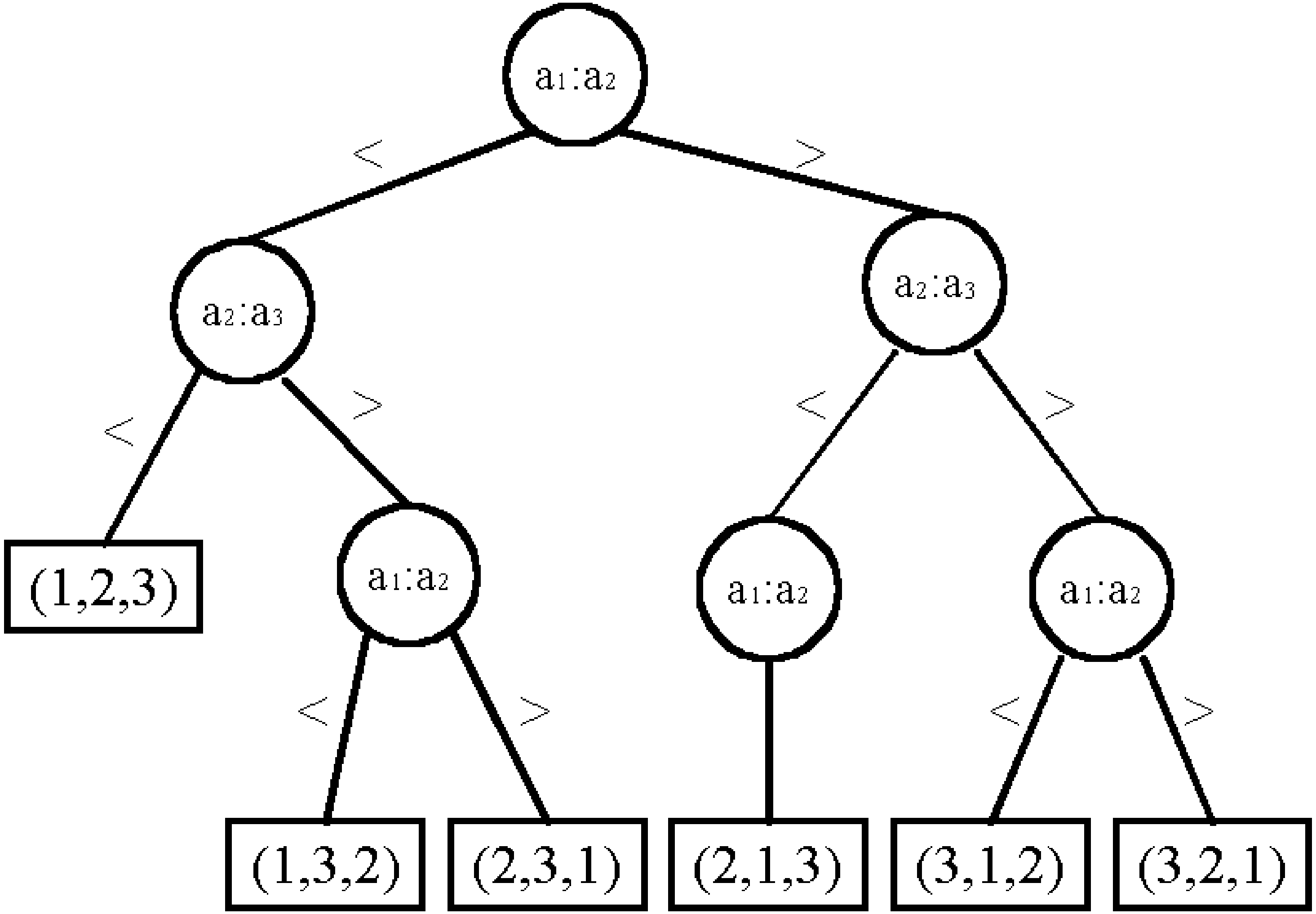
  (1)a_1:a_2, $a_1 \leftrightarrow a_2$

  (2)a_2:a_3

Decision tree for straight insertion sort



Decision tree for bubble sort



Lower bound of sorting

- The action of a sorting algorithm based upon compare and exchange operations on a particular input data set corresponds to one path from the top of the tree to a leaf node
- Each ***leaf node*** therefore corresponds to a particular permutation.
- *The longest path from the top of the tree to a leaf node, which is called the depth of the tree, represents the worst case time-complexity o this algorithm.*
- To find the lower bound of the sorting problem, we have to find the smallest depth of some tree, among all possible binary decision trees modeling sorting algorithms.

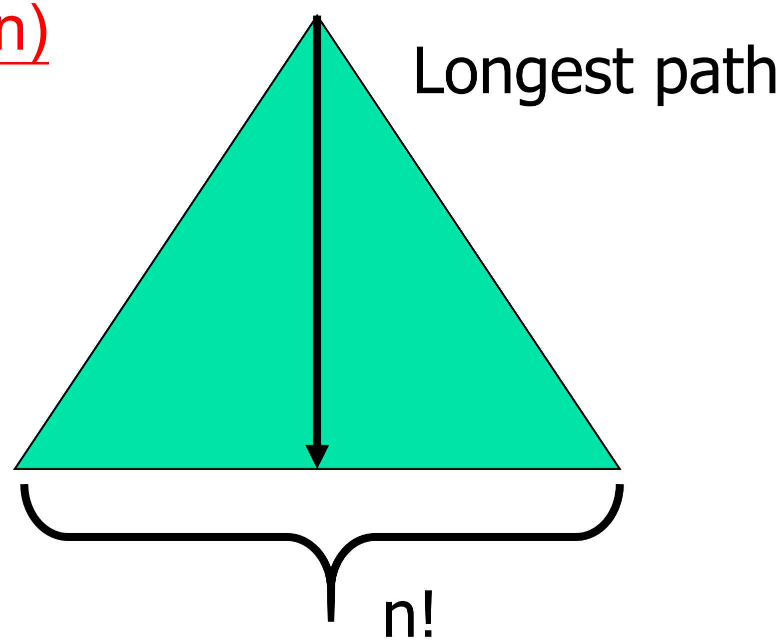
Lower bound of sorting

- To find the lower bound, we have to find the depth of a binary tree with the smallest depth.
- *n!* distinct permutations

  n! leaf nodes in the binary decision tree.
- balanced tree has the smallest depth:

  $\lceil \log(n!) \rceil = \Omega(n \log n)$

  lower bound for sorting: Ω(n log n)



Method 1:

$$\int \ln x dx = x \ln x - x + C$$

$$\int \log_\alpha x dx = \frac{1}{\ln \alpha}(x \ln x - x) + C$$

$$\log(n!) = \log(n(n-1)\cdots 1)$$

$$= \log 2 + \log 3 + \cdots + \log n = (2-1)\log 2 + (3-2)\log 3 + \ldots + (n-n+1)\log n$$

$$> \int_1^n \log x dx$$

$$= \log e \int_1^n \ln x dx$$

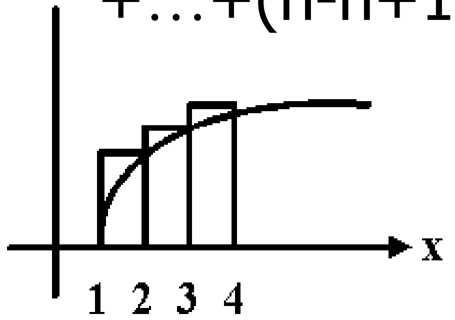
$$= \log e[x \ln x - x]_1^n$$

$$= \log e(n \ln n - n + 1)$$

$$= n \log n - n \log e + 1.44$$

$$\geq n \log n - 1.44n$$

$$= \Omega(n \log n)$$



Change base

$$\log_a b = \frac{\ln b}{\ln a}$$

Method 2:

- Stirling approximation:

$$\boxed{n! \approx S_n = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n}$$

$$\log n! \approx \log\sqrt{2\pi} + \frac{1}{2}\log n + n\log\frac{n}{e} \approx n\log n = \Omega(n\log n)$$

n	n!	S_n
1	1	0.922
2	2	1.919
3	6	5.825
4	24	23.447
5	120	118.02
6	720	707.39
10	3,628,800	3,598,600
20	2.433×10^{18}	2.423×10^{18}
100	9.333×10^{157}	9.328×10^{157}

Question:

- What is the highest lower bound of worst case time complexity of the comparison-based sorting algorithm?

(1) $\Omega(1)$

(2) $\Omega(n)$

(3) $\Omega(n \log n)$

(4) $\Omega(n^2)$.

Ans. 4

Knockout sort & Heap Sort

學習目標

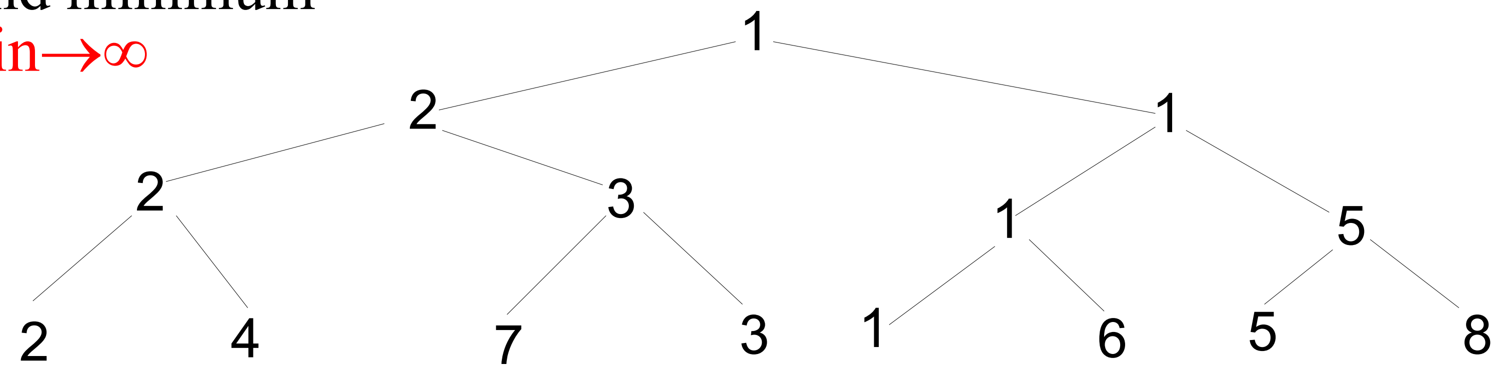
- Knockout 排序演算法設計
- Knockout 排序演算法的複雜度分析。
- 堆積排序 (heap sort) 演算法設計
- 堆積排序 (heap sort) 演算法的複雜度分析
  。

knockout sort

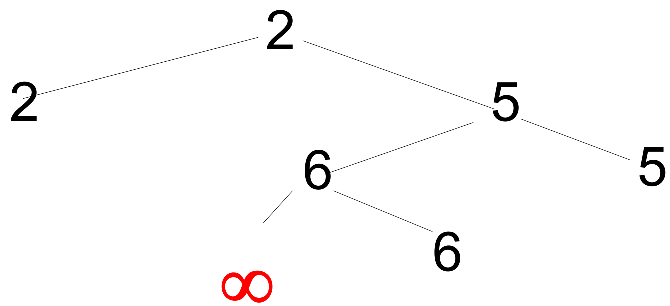
- Note that when we try to find the second smallest number, the information we may have extracted by finding the first smallest number is not used at all.
- This is why the straight insertion sort behaves so clumsily.
- It keeps some information after it finds the first smallest number so that it is quite efficient to find the <u>second smallest number</u>.

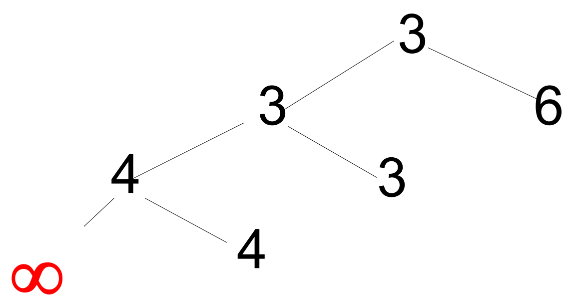
Knockout sort (example)

Input: 2, 4, 7, 3, 1, 6, 5,8
Construct Knockout tree
Find minimum
Min→∞



(n-1) comparisons





. . . . .

($\lceil \log n \rceil$ -1)

Time complexity of Knockout (淘汰) sort

- The first smallest number is found after $(n-1)$ comparisons.
- For all of the other selections, only $\lceil \log n \rceil - 1$ comparisons are needed. Therefore the total number of comparisons is

  $$(n-1)+(n-1)(\lceil \log n \rceil - 1).$$
- Thus the time-complexity of knockout sort is $O(n\log n)$ which is equal to the lower.
- Knockout sort is therefore <u>an optimal sorting</u> algorithm.
- We must note that the time-complexity $O(n\log n)$ is valid for best, average and worst cases.
- Drawbacks: **space 2n**.

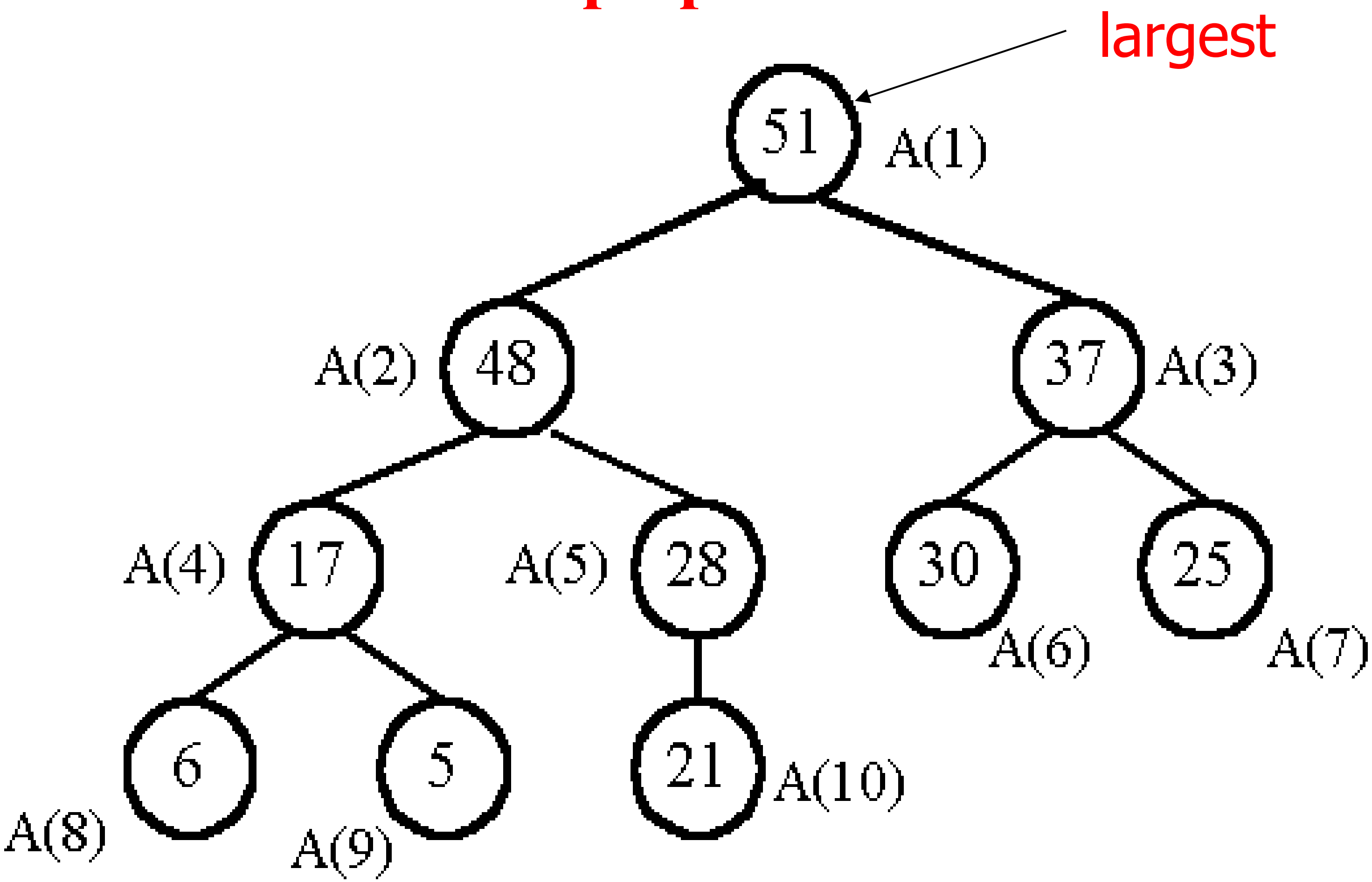
Heap

- A heap is a binary tree satisfying the following conditions:
  - This tree is completely balanced.
  - If the height of this binary tree is *h,* then leaves can be at level *h* or level h-1.
  - All leaves at level *h* are as far to the left as possible.
  - The data associated with all descendants of a node are smaller than the datum associated with this node.

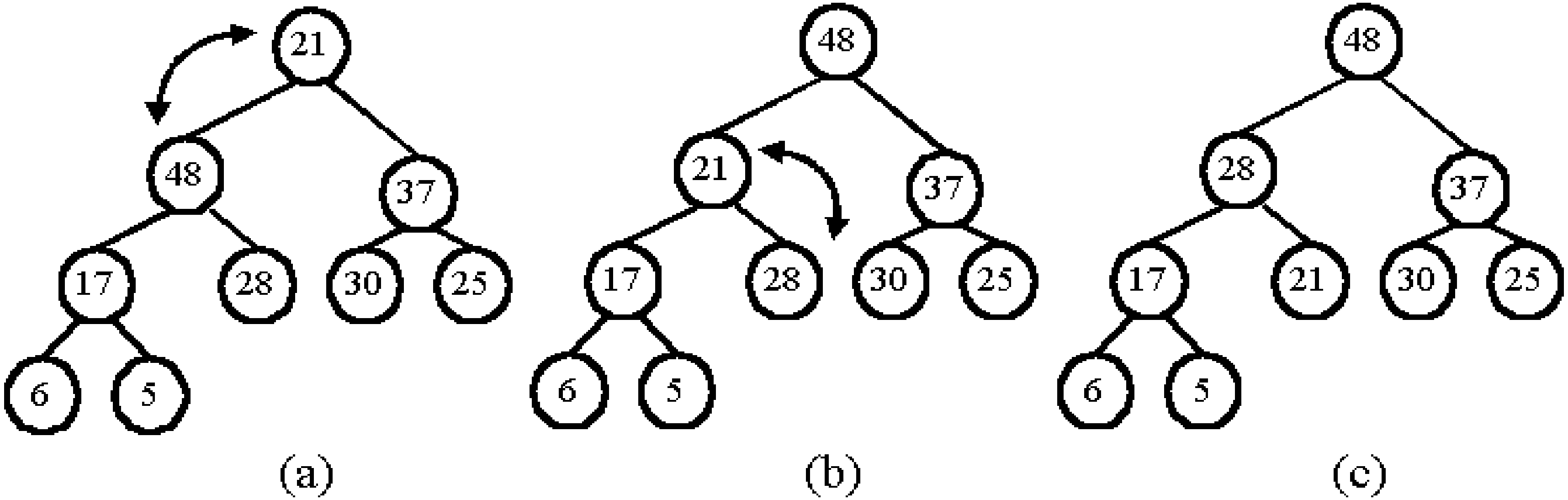
Max-heap or min-heap

Heapsort—An optimal sorting algorithm

- A **maximal heap : parent ≥ son**



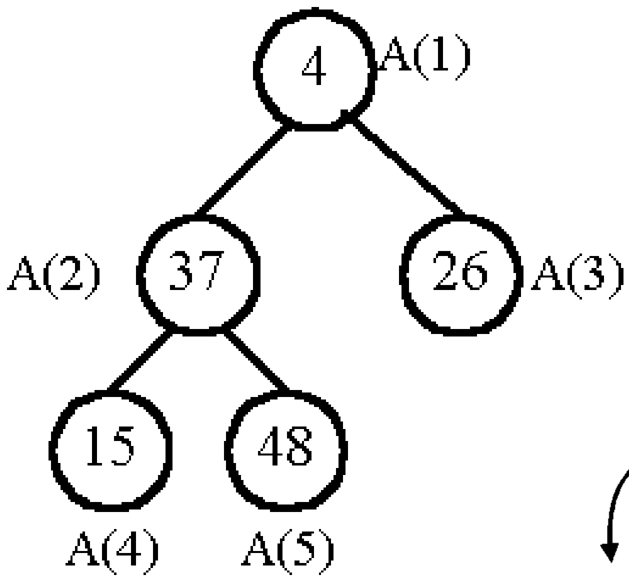
- output the maximum and restore:



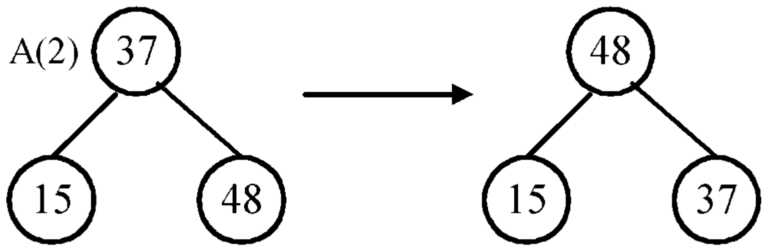
- Heapsort:
  - Phase 1: Construction
  - Phase 2: Output

Phase 1: construction

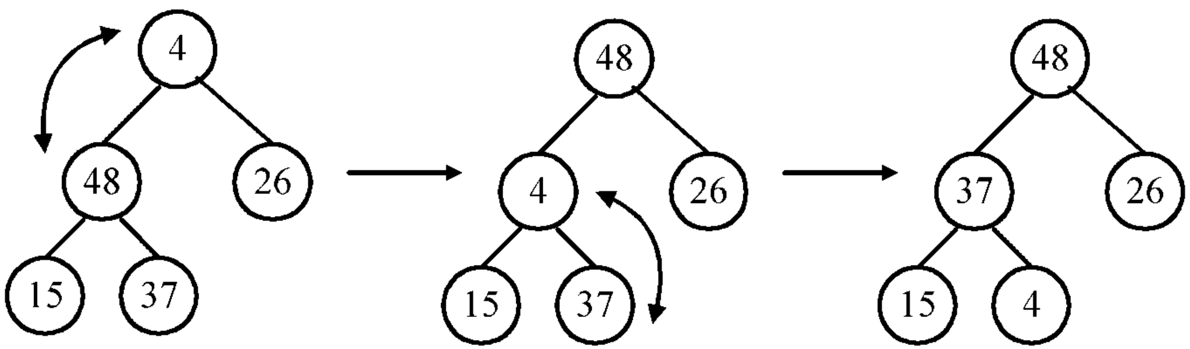
- input data: 4, 37, 26, 15, 48



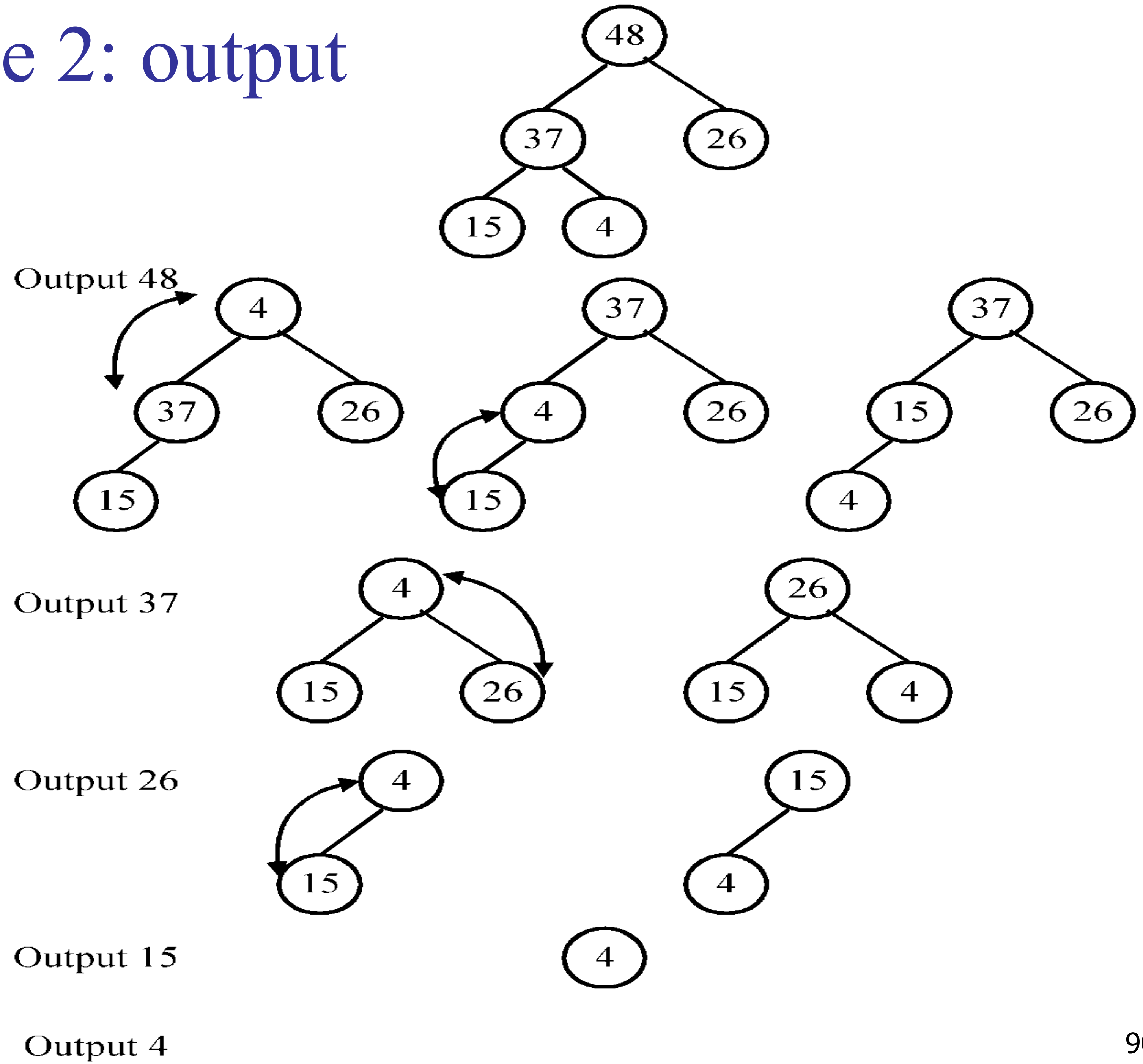
- restore the subtree rooted at A(2):



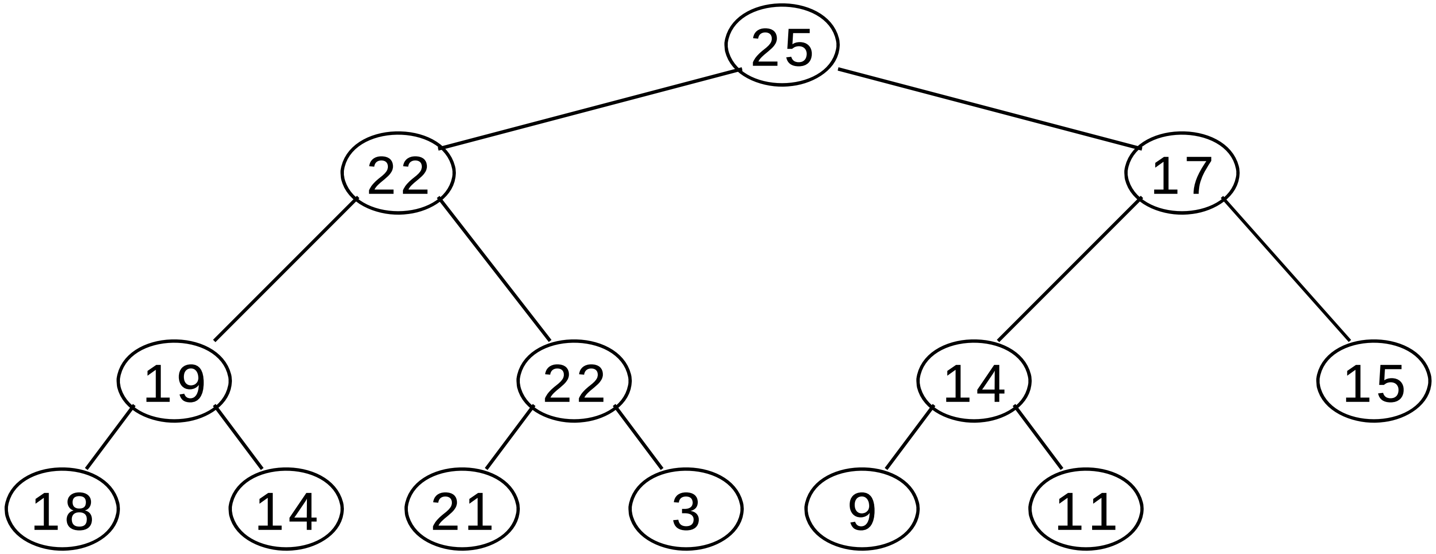
- restore the tree rooted at A(1):



Phase 2: output



Implementation heap sort



0	1	2	3	4	5	6	7	8	9	10	11	12
25	22	17	19	22	14	15	18	14	21	3	9	11

- Notice: (for iniital index as 0)
  - The left child of index i is at index 2*i+1
  - The right child of index i is at index 2*i+2
  - Example: the children of node 19 (3) are 18 (7) and 14 (8)

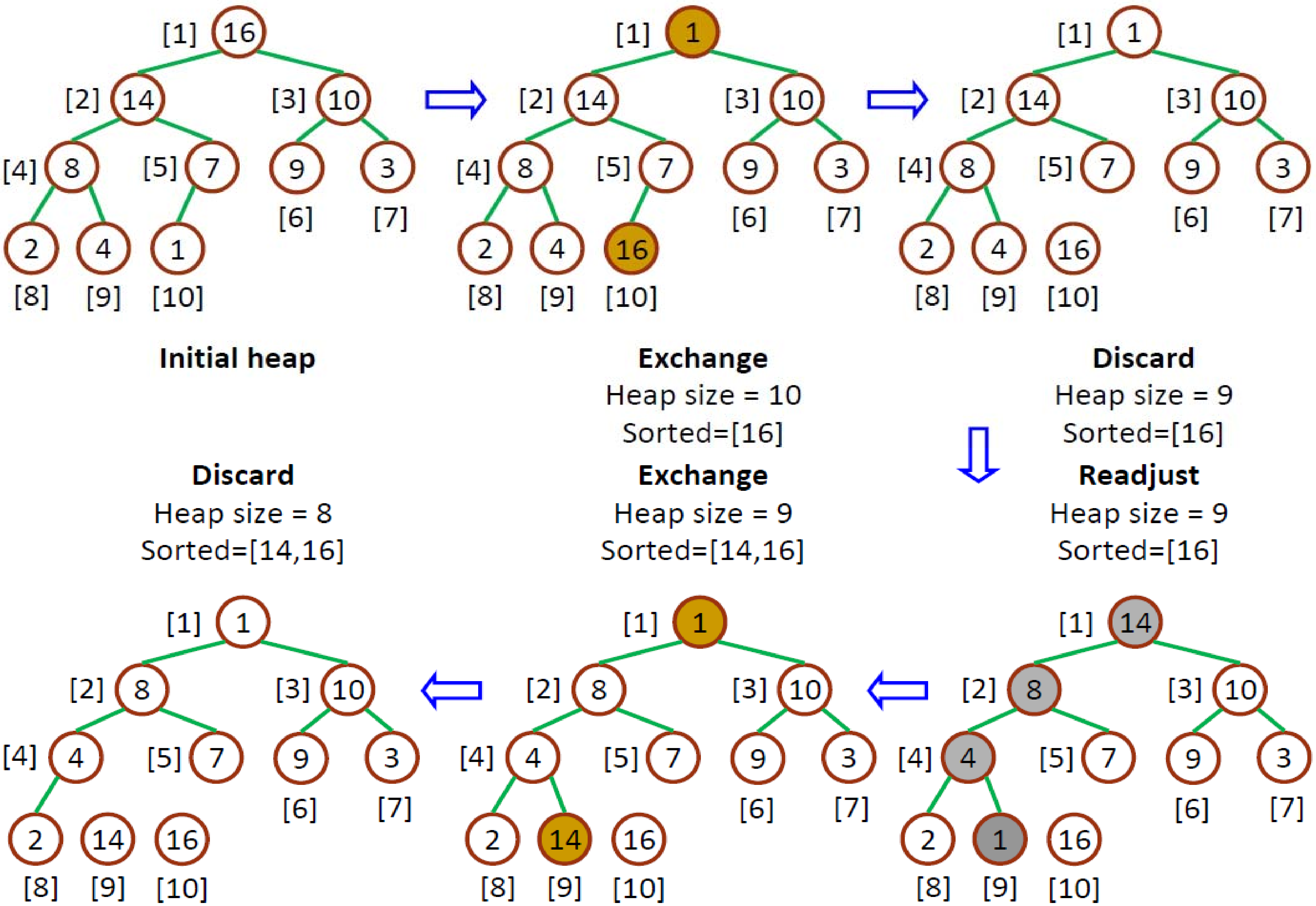
Removing and replacing the root

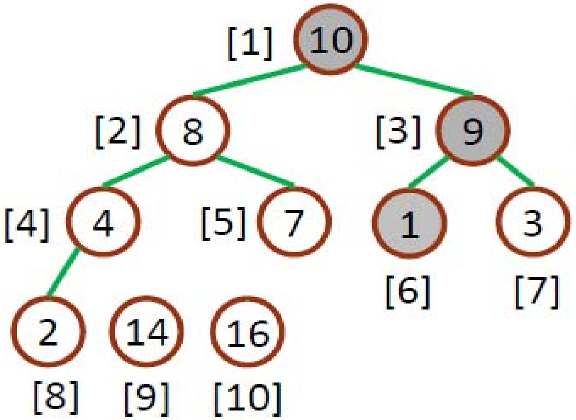
- The "root" is the first element in the array
- The "rightmost node at the deepest level" is the last element
- Swap them...

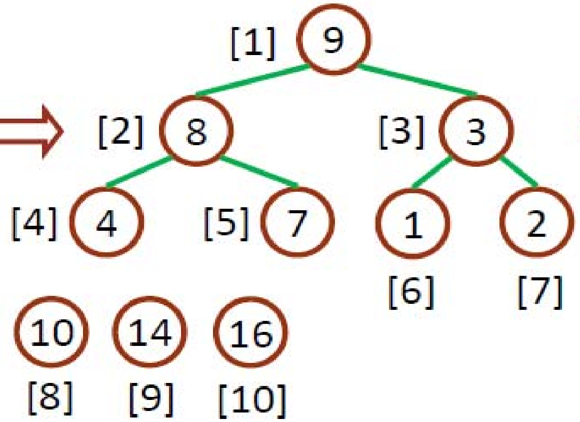
0	1	2	3	4	5	6	7	8	9	10	11	12
25	22	17	19	22	14	15	18	14	21	3	9	11

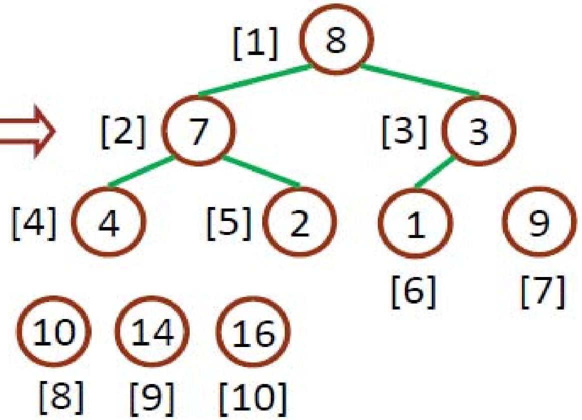
0	1	2	3	4	5	6	7	8	9	10	11	12
11	22	17	19	22	14	15	18	14	21	3	9	25

- ...And pretend that the last element in the array no longer exists—that is, the "last index" is 11 (9)


Initial heap
Exchange
Heap size = 10
Sorted=[16]
Discard
Heap size = 9
Sorted=[16]
Readjust
Heap size = 9
Sorted=[16]
Exchange
Heap size = 9
Sorted=[14,16]
Discard
Heap size = 8
Sorted=[14,16]







Readjust
Heap size = 8
Sorted=[14,16]

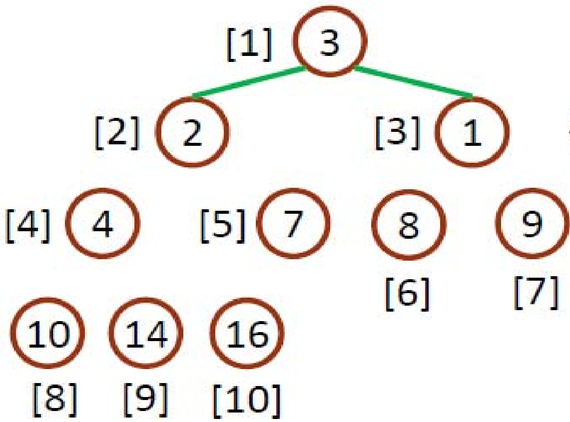
Heap size = 7
Sorted=[10,14,16]

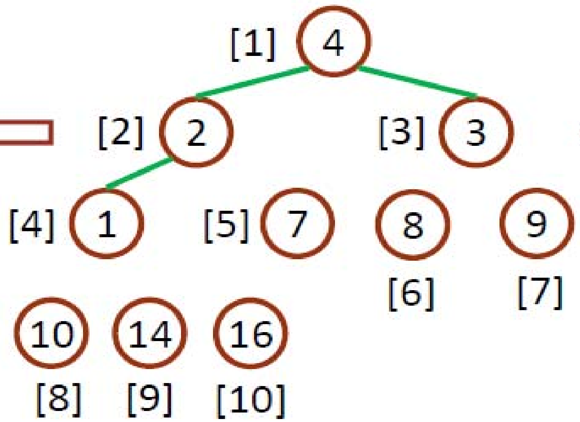
Heap size = 6
Sorted=[9,10,14,16]

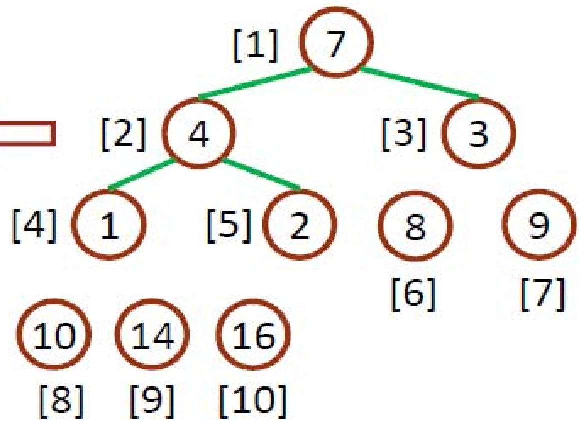
Heap size = 3
Sorted=[4,7,8,9,10,14,16]

Heap size = 4
Sorted=[7,8,9,10,14,16]

Heap size = 5
Sorted=[8,9,10,14,16]







Time complexity Phase 1: construction

Let the level of an internal node be L. The worst case **2(d-L)** comparisons Have to be made to perform the restore.

$\boxed{d = \lfloor \log n \rfloor : \text{depth}}$

# of comparisons is at most:

2^L: number of nodes in level L

$$\sum_{L=0}^{d-1} 2(d-L)2^L$$

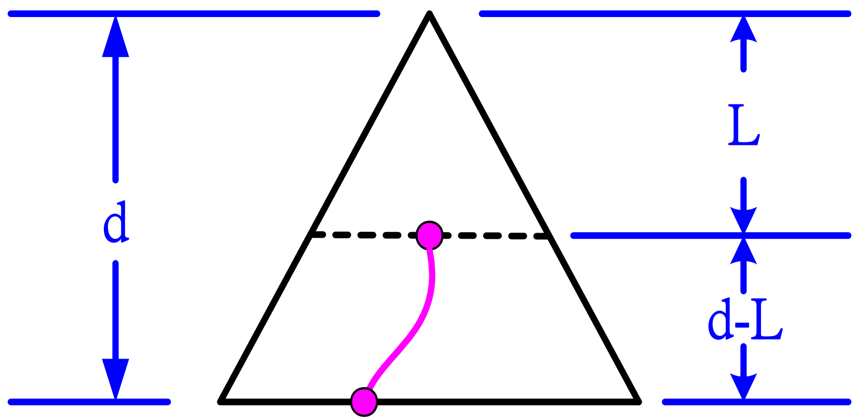
$$= 2d \sum_{L=0}^{d-1} 2^L - 4 \sum_{L=0}^{d-1} L2^{L-1}$$

$$\boxed{\left(\sum_{L=0}^{k} L2^{L-1} = 2^k(k-1)+1 \right)}$$

$$= 2d(2^d-1) - 4(2^{d-1}(d-1-1)+1)$$

$$\vdots$$

$$= cn - 2\lfloor \log n \rfloor - 4, \quad 2 \le c \le 4$$



Time complexity

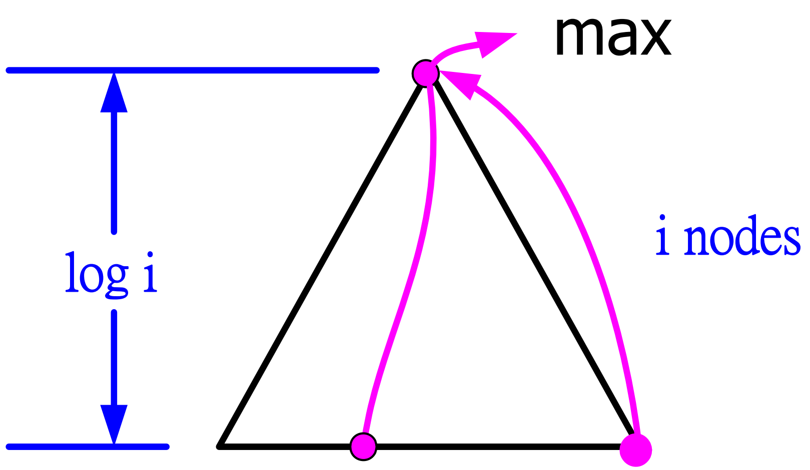
Phase 2: output (delete element from heap)

$$2\sum_{i=1}^{n-1} \lfloor \log i \rfloor$$

$$= \quad \vdots$$

$$= 2n\lfloor \log n \rfloor - 4cn + 4, \quad 2 \le c \le 4$$

$$= O(n \log n)$$



$$2\sum_{i=1}^{n-1}\lfloor\log i\rfloor$$

For n=10 we have $\lfloor\log 1\rfloor = 0$

$\lfloor\log 2\rfloor$=$\lfloor\log 3\rfloor$=1

$\lfloor\log 4\rfloor$=$\lfloor\log 5\rfloor = \lfloor\log 6\rfloor$=$\lfloor\log 7\rfloor$=2

$\lfloor\log 8\rfloor$=$\lfloor\log 9\rfloor$=3

Thus,

there are 2^1 numbers equal to $\lfloor\log 2\rfloor$=1

there are 2^2 numbers equal to $\lfloor\log 2^2\rfloor$=2

And

$$\boxed{10 - 2^{\lfloor\log 10\rfloor}=10-2^3=2 \text{ numbers equal to } \lfloor\log n\rfloor}$$

$$2\sum_{i=1}^{n-1}\lfloor \log i \rfloor$$

$$= \boxed{2\sum_{i=1}^{\lfloor \log n \rfloor -1} i2^{i}} + 2(n - 2^{\lfloor \log n \rfloor})\lfloor \log n \rfloor$$

$$= 4\sum_{i=1}^{\lfloor \log n \rfloor -1} i2^{i-1} + 2(n - 2^{\lfloor \log n \rfloor})\lfloor \log n \rfloor.$$

Using $\boxed{\sum_{i=1}^{k} i2^{i-1} = 2^{k}(k-1)+1}$ (

$$2\sum_{i=1}^{n-1}\lfloor \log i \rfloor$$

$$= 4\sum_{i-1}^{\lfloor \log n \rfloor-1} i2^{i-1} + 2(n - 2^{\lfloor \log n \rfloor})\lfloor \log n \rfloor$$

$$= 4(2^{\lfloor \log n \rfloor-1}(\lfloor \log n \rfloor - 1 - 1) + 1) + 2n\lfloor \log n \rfloor - 2\lfloor \log n \rfloor 2^{\lfloor \log n \rfloor}$$

$$= 2 \cdot 2^{\lfloor \log n \rfloor}\lfloor \log n \rfloor - 8 \cdot 2^{\lfloor \log n \rfloor-1} + 4 + 2n\lfloor \log n \rfloor - 2 \cdot 2^{\lfloor \log n \rfloor}\lfloor \log n \rfloor$$

$$= 2 \cdot n\lfloor \log n \rfloor - 4 \cdot 2^{\lfloor \log n \rfloor} + 4$$

$$= 2n\lfloor \log n \rfloor - 4cn + 4 \qquad \text{where} \quad 2 \le c \le 4$$

$$= O(n \log n).$$

Question:

- In a minimal heap H[0..n-1], the left child of the element at index i is at index ?

(1) 2i+1

(2) 2i

(3) 2i-1

(4) 2i+2.

Ans. 1

Average case lower bound of sorting

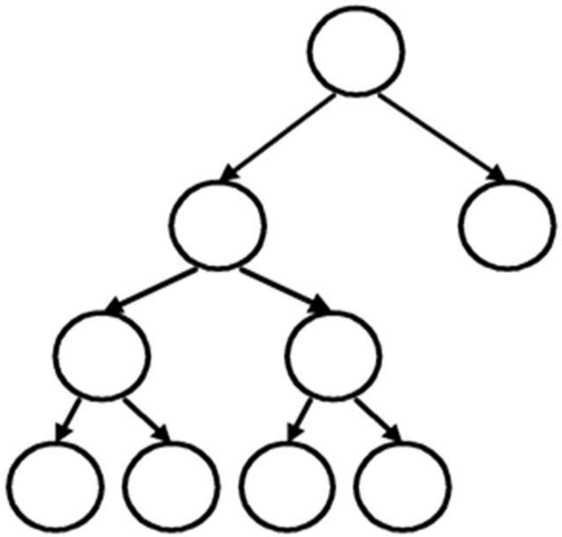
學習目標

- 排序演算法average case lower bound 的複雜度分析。

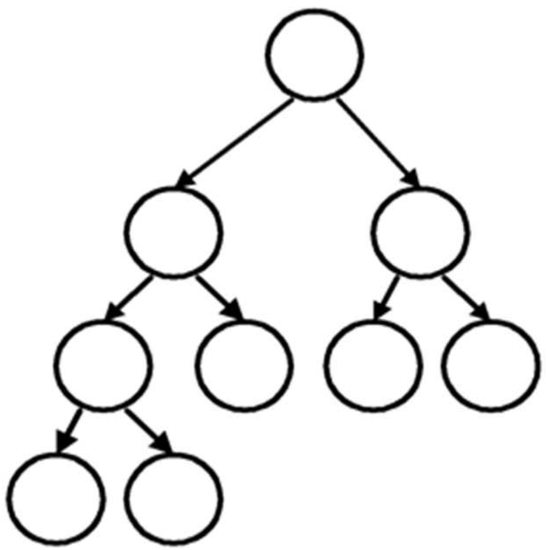
Average case lower bound of sorting

- By binary decision tree
- The length of this path is equal to the number of comparisons executed for this input data set.
- The average time complexity of a sorting algorithm:
  - the **external path length** of the binary tree is the sum of the lengths of paths from root to each leaf node.
  - Leaf number： n!
- The external path length is minimized if the tree is balanced.

  (all leaf nodes on level d or level d–1)

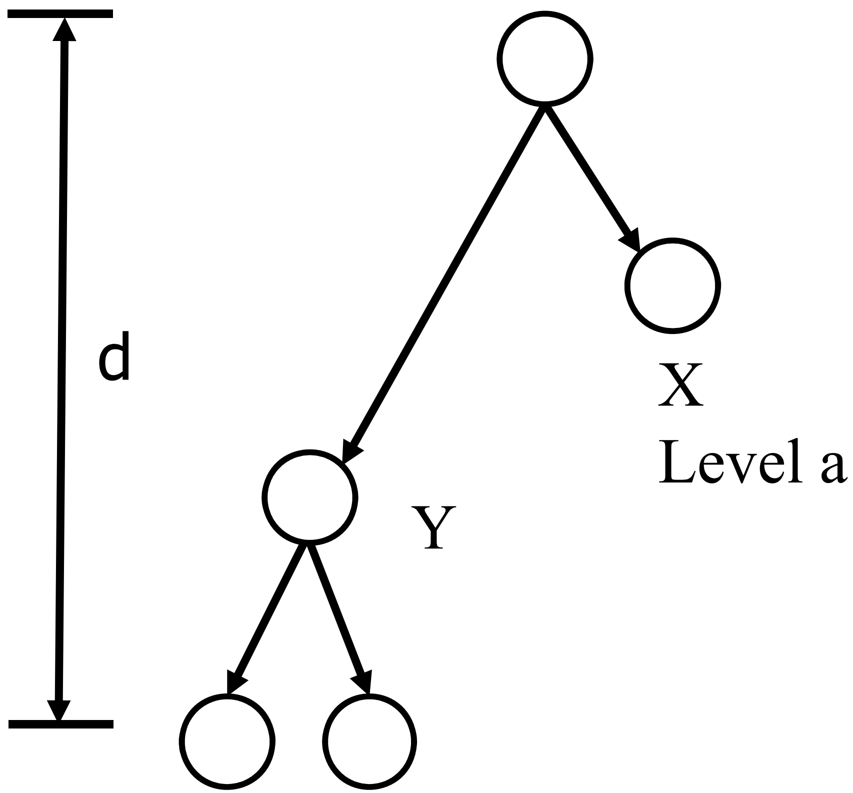


unbalanced
external path length
$= 4 \cdot 3 + 1 = 13$



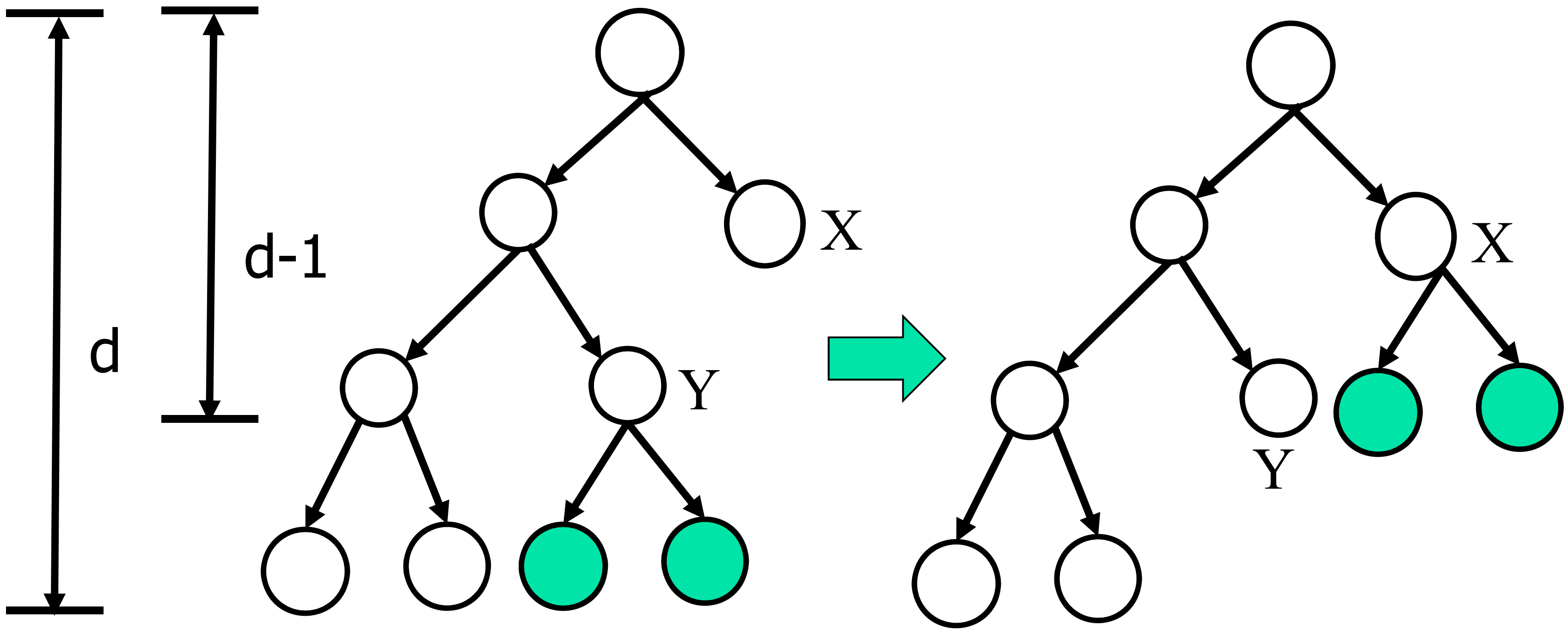
balanced
external path length
$= 2 \cdot 3 + 3 \cdot 2 = 12$

Tree Modification



- Modify the tree such the external path length is decreased without changing the # of leaf nodes.

The tree can be modified such that the external path length is decreased without changing the number of leaf node.



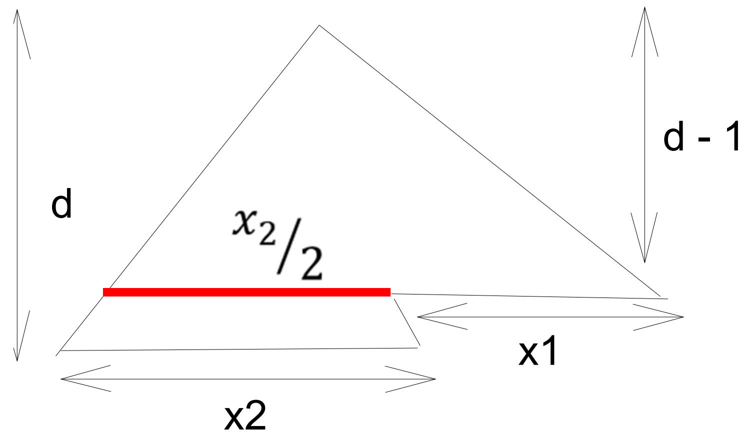
The external path length of a binary tree is minimized if and only if the tree is balanced.

Compute the min external path length

1. Depth of balanced binary tree with c leaf nodes:
   depth $d = \lceil \log c \rceil$
   Leaf nodes can appear only on level d or d–1(balanced).
2. x_1 leaf nodes on level d–1
   x_2 leaf nodes on level d
   - $x_1 + x_2 = c$
   - $x_1 + \frac{x_2}{2} = 2^{d-1}$

   $\Rightarrow \quad x_1 = 2^d - c$
   $\quad\quad x_2 = 2(c - 2^{d-1})$

•Assume x_2 is even.
•Two leave in level d has a parent in level d-1



The external path length of a balanced binary tree is the lower bound of the sorting(in average case).

3. External path length:

$$M = x_1(d-1) + x_2 d$$
$$= (2^d - c)(d-1) + 2(c - 2^{d-1})d$$
$$= c + cd - 2^d, \qquad \log c \le d < \log c + 1$$
$$\ge c + c\log c - 2 * 2^{\log c}$$
$$= \mathbf{c \log c - c}$$

4. $c = n!$

$$M = n! \log n! - n!$$
$$M/n! = \log n! - 1$$
$$= \Omega(n \log n)$$

Average case lower bound of sorting: $\Omega(n\log n)$

Quicksort & Heapsort

- Quicksort is optimal in the average case.
  (O(n log n) in average)
- (i) worst case time complexity of heapsort is
  O(n log n)
  (ii) average case lower bound: Ω(n log n)
  - average case time complexity of heapsort is O(n log n)
  - Heapsort is optimal in the average case.

Question:

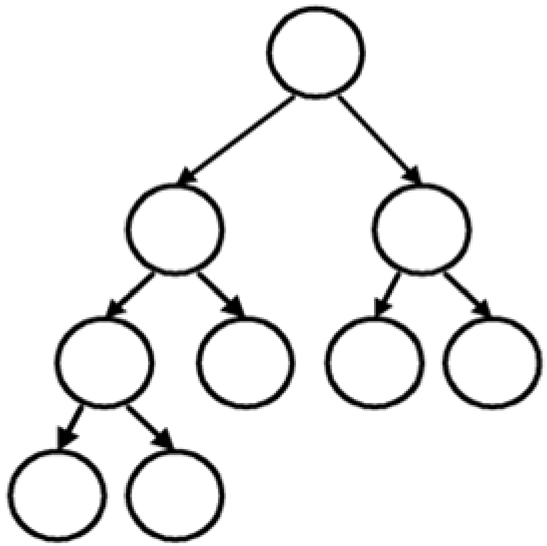
- For the given binary tree, hat is the value of the external path length?

(1) 5

(2) 9

(3) 12

(4) 13.



Ans. 3

Improving a lower bound through oracles

學習目標

- Lower Bound (LB) 證明技巧
  - Oracle 法
  - Problem Transformation 法

Improving a lower bound through oracles

- In some cases, the binary decision tree model does not produce a very meaningful LB. (can be improved)
- Problem P: merge two sorted sequences A and B with lengths m and n.
- Conventional 2-way merging:

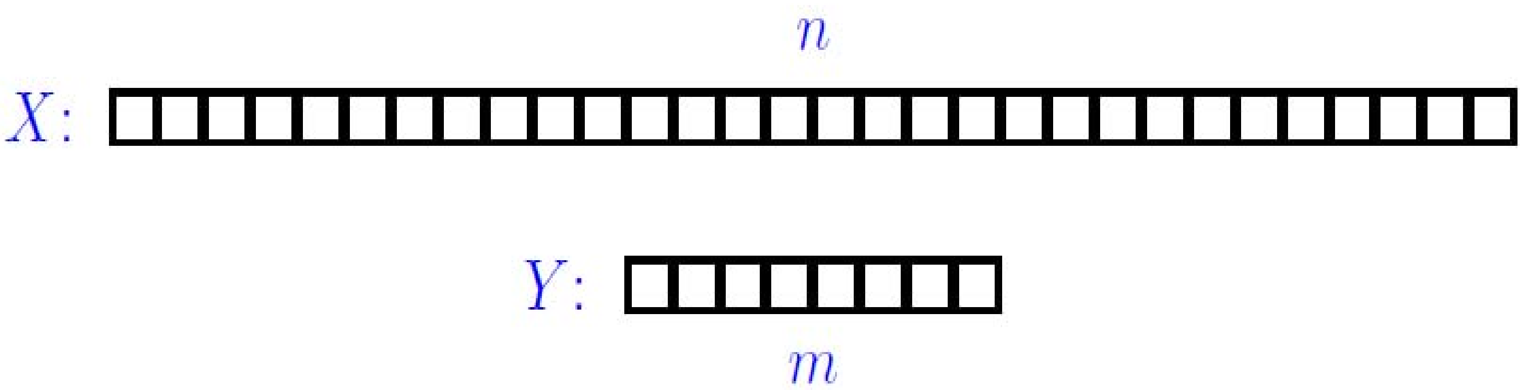
2 3 5 6

1 4 7 8

- Complexity: at most m+n-1 comparisons

Input: Two sorted lists X and Y of length n and m.

We may assume $n \geq m$.



Standard Merge:

$$\Theta(n+m)$$

Binary Insertion of Y in X:

$$\Theta(m \log n))$$

(1) Binary decision tree:

- **How many possible different merged sequence are there?**
- **Assume (m+n) elements are distinct.**

There are $\binom{m+n}{n}$ ways to merge n elements into m elements without disturbing the original order. (why?)

$\binom{m+n}{n}$ leaf nodes in the decision tree (all possible cases).

$\Rightarrow$ The lower bound for merging:

$$\left\lceil \log \binom{m+n}{n} \right\rceil \leq m + n - 1$$

(conventional merging)

- When m = n

$$\log\binom{m+n}{n} = \log\frac{(2m)!}{(m!)^2} = \log((2m)!) - 2\log m!$$

Using Stirling approximation

$$n! \approx \sqrt{2\pi n}\,(\frac{n}{e})^n$$

$$\log\binom{m+n}{n} \approx (\log\sqrt{2\pi} + \log\sqrt{2m} + 2m\log\frac{2m}{e}) -$$

$$-2\,(\log\sqrt{2\pi} + \log\sqrt{m} + m\log\frac{m}{e})$$

$$\approx 2m - \frac{1}{2}\log m + O(1) \;<\; 2m-1$$

- Optimal algorithm: conventional merging needs 2m-1 comparisons

(2) Oracle (聖賢;哲人):

- The oracle tries its best to cause the algorithm to work as hard as it might. (to give a very hard data set)->to find worst case.
- Two sorted sequences:
  - A: $a_1 < a_2 < \ldots < a_m$
  - B: $b_1 < b_2 < \ldots < b_m$
- The very hard case:
  - $\mathbf{a_1 < b_1 < a_2 < b_2 < \ldots < a_m < b_m}$

- We must compare:

$$\begin{aligned} &a_1 : b_1 \\ &b_1 : a_2 \\ &a_2 : b_2 \\ &\ \vdots \\ &b_{m-1} : a_{m-1} \\ &a_m : b_m \end{aligned}$$

- Otherwise, we may get a wrong result for some input data.

  e.g. If b_1 and a_2 are not compared, we can not distinguish

  $a_1 < b_1 < a_2 < b_2 < \ldots < a_m < b_m$ and

  $a_1 < a_2 < b_1 < b_2 < \ldots < a_m < b_m$
- Thus, at least 2m–1 comparisons are required.
- The conventional merging algorithm is optimal for m = n.

Finding lower bound by problem transformation

Finding lower bound by problem transformation

Problem A reduces to problem B ($A \propto B$)

iff A can be solved by using any algorithm which solves B.

If $A \propto B$, B is more difficult.

$$
\begin{tikzcd}[column sep=huge, row sep=large]
\text{instance of } A \arrow[r, "\text{transformation}", "T(tr_1)"'] \arrow[d, "T(A)"'] & \text{instance of } B \arrow[d, "T(B)\quad \text{solver of } B"] \\
\text{answer of } A & \text{answer of } B \arrow[l, "\text{transformation}"', "T(tr_2)"]
\end{tikzcd}
$$

Note: $T(tr1) + T(tr2) < T(B)$

$T(A) \leq T(tr1) + T(tr2) + T(B) \sim O(T(B))$

The lower bound of the convex hull problem

- sorting $\propto$ convex hull

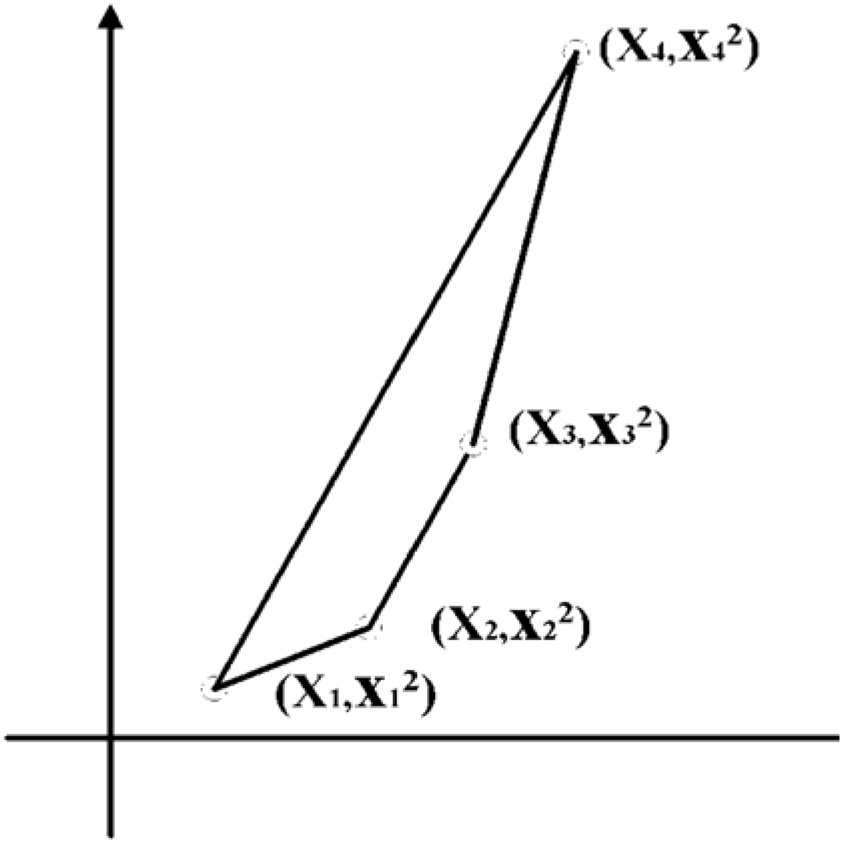
  A B

- an instance of A: $(x_1, x_2, \ldots, x_n)$

  ↓transformation

  an instance of B: $\{(x_1, x_1^2), (x_2, x_2^2), \ldots, (x_n, x_n^2)\}$

  assume: $x_1 < x_2 < \ldots < x_n$



Solve A by transform A to B, and solve B, the result of B can be Easily transformed to the solution of A.

Reduction to convex hull problem

- The reduction of Sorting problem to Convex Hull problem:
  - Reduction **sortByConvexHull(S)**
  - {// S is a sequence of numbers.
    - 1. for i in 1..n, set P[i] = point(S[i], $S[i]^2$);

    /* in other words, set P = { (x, x^2) | x in S } */

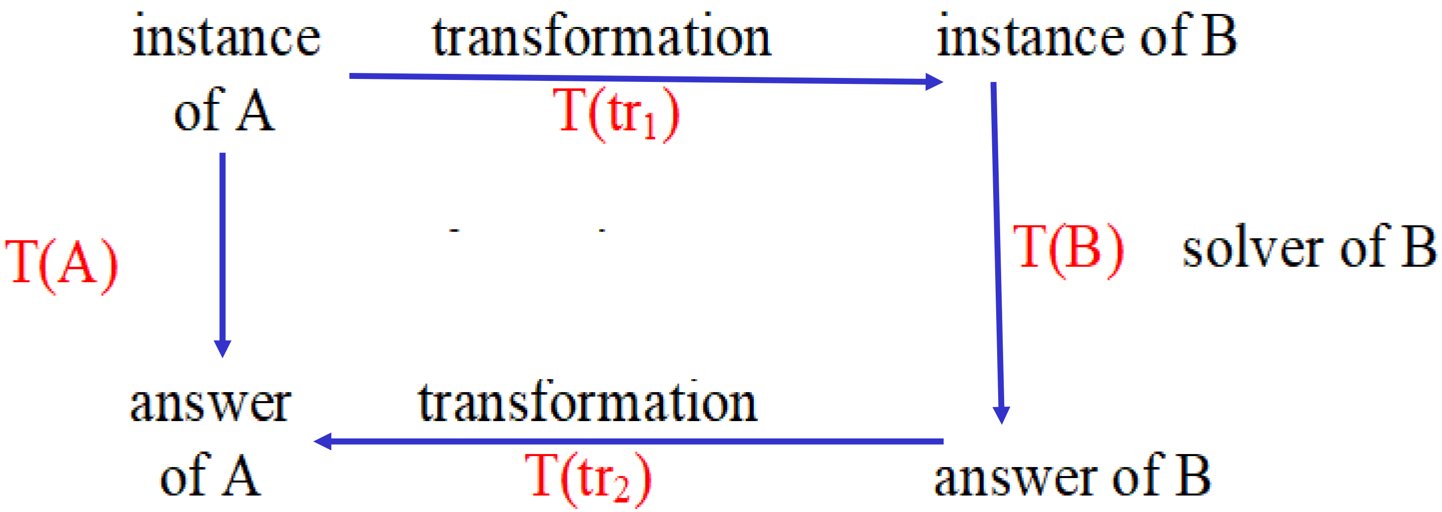
    - 2. k = **convexHull(P)**;

    /* We know in advance that k will be size(P).*/

    - 3. find the point with smallest x,
    - for i in 1..n, set S[i] = **P[i].first**;

    /* first = the x of a (x, x^2) pair. */
  - }

- If the convex hull problem can be solved, we can also solve the sorting problem.
    - The lower bound of sorting: $\Omega(n \log n)$
- <u>The lower bound of the convex hull problem: $\Omega(n \log n)$</u>



The lower bound of the Euclidean minimal spanning tree (MST) problem

- sorting $\propto$ Euclidean MST

  A B

- an instance of A: $(x_1, x_2, \ldots, x_n)$

  $\downarrow$transformation

  an instance of B: $\{(x_1, 0), (x_2, 0), \ldots, (x_n, 0)\}$

  - Assume $x_1 < x_2 < x_3 < \ldots < x_n$
  - $\Leftrightarrow$there is an edge between $(x_i, 0)$ and $(x_{i+1}, 0)$ in the MST, where $1 \leq i \leq n-1$

- If the Euclidean MST problem can be solved, we can also solve the sorting problem.
  - The lower bound of sorting: $\Omega(n \log n)$
- <u>The lower bound of the Euclidean MST problem: $\Omega(n \log n)$</u>

Sorting In Linear Time

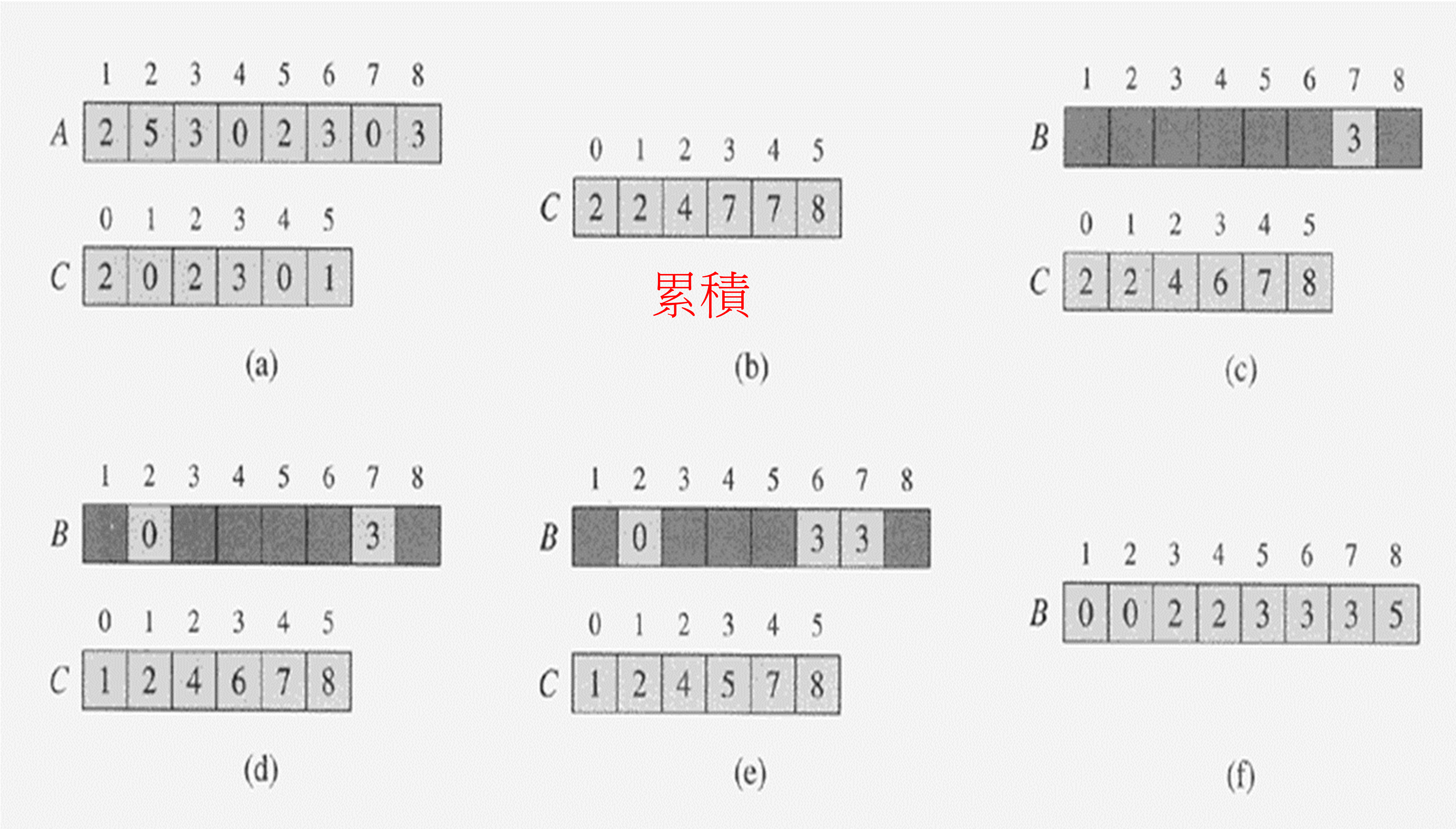
學習目標

- 線性時間排序演算法
  - Counting sort 法
  - Radix sort 法
  - Bucket sort 法

Sorting In Linear Time

- Counting sort
  - No comparisons between elements!
  - ***But***...depends on assumption about the numbers being sorted
    - We assume numbers are in the range ***1.. k***
  - The algorithm:
    - Input: A[1..n], where A[j] $\in$ {1, 2, 3, …, k}
    - Output: B[1..n], sorted (notice: not sorting in place)
    - Also: Array C[1..k] for auxiliary storage

Counting sort



Counting Sort

```
CountingSort(A, B, k)
    for i=1 to k
        C[i]= 0;
    for j=1 to n
        C[A[j]] += 1;
    for i=2 to k
        C[i] = C[i] + C[i-1];
    for j=n downto 1
        B[C[A[j]]] = A[j];
        C[A[j]] -= 1;
```

Takes time $O(k)$

Takes time $O(n)$

What will be the running time?

Counting Sort

- Total time: $O(n + k)$
  - Usually, $k = O(n)$
  - Thus counting sort runs in $O(n)$ time
- But sorting is $\Omega(n \log n)$!
  - No contradiction--this is not a comparison sort (in fact, there are *no* comparisons at all!)
  - Notice that this algorithm is *stable*

穩定排序法(stable sorting)，如果鍵值相同之資料，在排序後相對位置與排序前**相同**時，稱**穩定排序**。

【例如】

排序前：3,5,19,1,3*,10

排序後：1,3,3*,5,10,19

(因為兩個3, 3*的相對位置在排序前與後皆相同。)

Counting Sort

- Cool! *Why don't we always use counting sort?*
- Because it depends on range k of elements
- *Could we use counting sort to sort 32 bit integers? Why or why not?*
- Answer: no, k too large ($2^{32} = 4{,}294{,}967{,}296$)

Counting Sort

- *How did IBM get rich originally?*
- Answer: punched card readers for census tabulation in early 1900's.
  - In particular, a *card sorter* that could sort cards into different bins
    - Each column can be punched in 12 places
    - Decimal digits use 10 places
  - Problem: only one column can be sorted on at a time

Radix sort

329	720	720	329
457	355	329	355
657	436	436	436
839	457	839	457
436	657	355	657
720	329	457	720
355	839	657	839

先排個位數

再排十位數

最後排百位數

Radix Sort

- Problem: lots of intermediate piles of cards (read: scratch arrays) to keep track of
- Key idea: sort the *least* significant digit first

```
RadixSort(A, d)
  for i=1 to d
    StableSort(A) on digit i
```

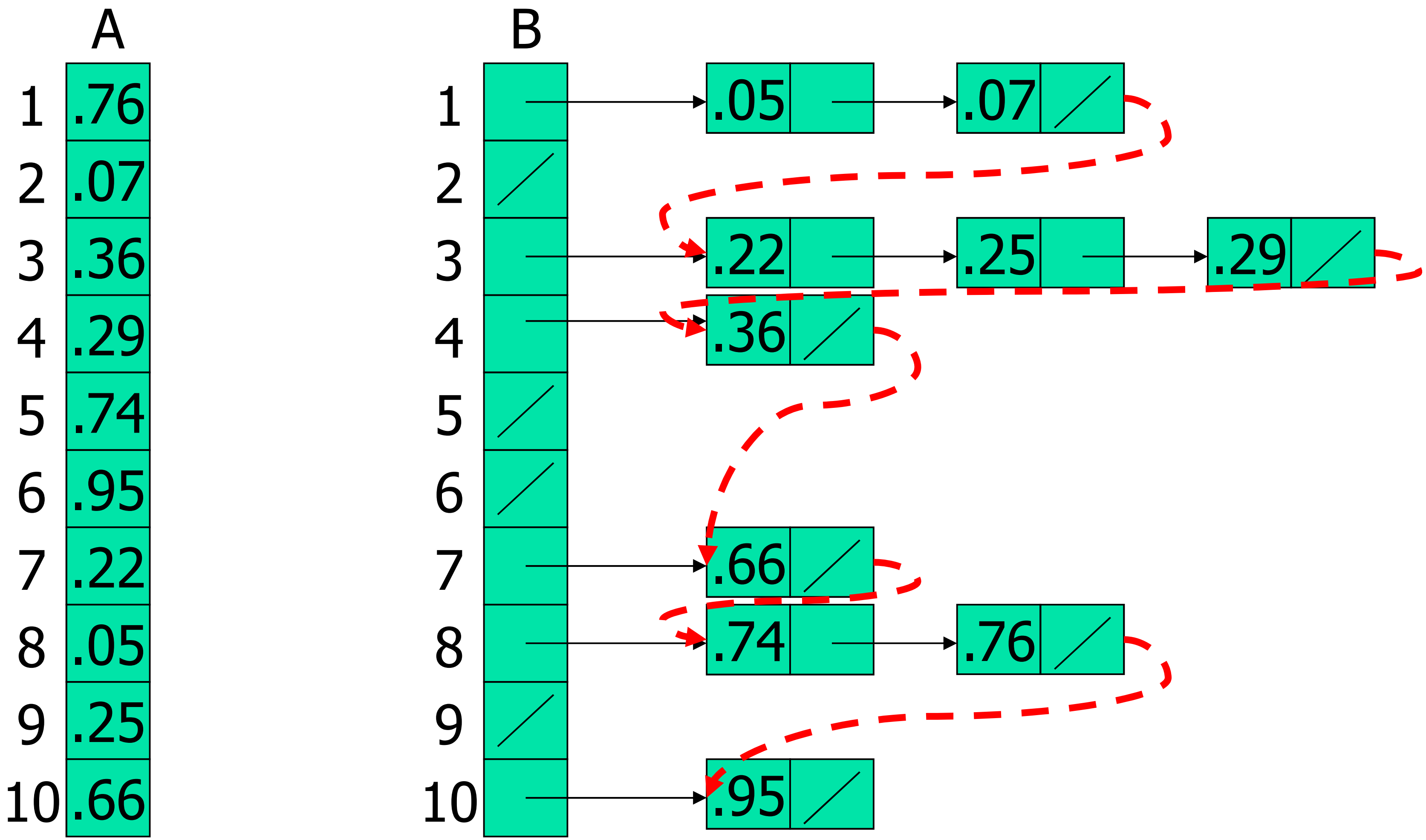
Radix Sort

- *What sort will we use to sort on digits?*
- Counting sort is obvious choice:
  - Sort n numbers on digits that range from $1..k$
  - Time: $O(n + k)$
- Each pass over n numbers with d digits takes time $O(n+k)$, so total time $O(dn+dk)$
  - When d is constant and $k=O(n)$, takes $O(n)$ time

Bucket Sort

- Bucket sort
  - Assumption: input is *n* reals from [0, 1)
  - Basic idea:
    - Create *n* linked lists (*buckets*) to divide interval [0,1) into subintervals of size 1/*n*
    - Add each input element to appropriate bucket and sort buckets with insertion sort
  - Uniform input distribution → O(1) bucket size
    - Therefore the expected total time is O(n)
  - These ideas will return when we study *hash tables*

Bucket Sort



A
B
1
2
3
4
5
6
7
8
9
10
.76
.07
.36
.29
.74
.95
.22
.05
.25
.66
1
2
3
4
5
6
7
8
9
10
.05
.07
.22
.25
.29
.36
.66
.74
.76
.95

Bucket Sort

```
BUCKET-SORT(A)
n ← length[A]
for i ← 1 to n
        do insert A[i] into list B[⌊n A[i]⌋]
for i ← 0 to n − 1
        do sort list B[i] with insertion sort
concatenate the lists B[0], B[1], . . . , B[n − 1] together in order
```