COMP251: Background

Jérôme Waldispühl
School of Computer Science
McGill University
Big-Oh notation

• “f(n) is $O(g(n))$” iff there exists a point $n_0$ beyond which $f(n)$ is less than some fixed constant times $g(n)$

For all $n \geq n_0$

$$f(n) \leq c \cdot g(n) \quad (\text{for } c = 1)$$

From COMP250 M. Blanchette’s slides
Running-time \( O(m+n) \) is equivalent to:

A. \( O(m) + O(n) \)
B. \( O(\max(m,n)) \)

Which of these possibilities are true?

- A and B
- A only
- B only
- neither A or B

45.2% of respondents chose A and B, 23.8% chose A only, 29.8% chose B only, and 0.2% chose neither A or B.
Let $T_1(n) = O(f(n))$ and $T_2(n) = O(f(n))$. Given the statements:

A. $T_1(n) / T_2(n) = O(1)$
B. $T_1(n) + T_2(n) = O(f(n))$

Which of them are true?

- A and B
- A only
- B only
- neither A or B

The pie chart shows the distribution of answers:
- 54.9% for B only
- 29.3% for A and B
- 8.5% for A only
- 5.3% for neither A or B
What is the time-complexity of the following piece of code in Big-Oh notation?

```cpp
sum = 0;
for (int i = 0; i < n; i++) {
    for (j = 1; j < n; j = j*2) {
        sum += n;
    }
}
```

- $O(n)$
- $O(n \cdot \log(n))$
- $O(n^2)$
- $O(\log(n))$

The pie chart indicates that $O(n \cdot \log(n))$ is the majority choice, with 24.1% of the votes.
for (i=1; i<N; i=i*2) { ... }

Value of i after k iterations: $2^k$

We have $i < N \Rightarrow 2^k < N \Rightarrow k < \log_2(N)$.

There is less than $\log_2(N)$ iterations, and the running time of this loop is $O(\log(n))$. 
For this Binary Tree, which of the following represents a post-order traversal?

- A, B, C, D, E, F, G, H, I
- F, B, A, D, C, E, G, I, H
- A, C, E, D, B, H, I, G, F
- None of the above

The correct answer is highlighted in red: A, C, E, D, B, H, I, G, F.
Post-order traversal

postorderTraversal(treeNode x)
    for each c in children(x) do
        postorderTraversal(c);
    print x.value;
For the Binary Search Tree shown above, deletion of node F would result in which of the following nodes becoming the root node?

- B
- G
- B or G
- E or G

E or G
BST - remove

1) Find the node N to be removed using the “find” algo
2) If (N is a leaf) { remove it }
   Else if (N is an internal node with only one child) {
       replace N by its child
   }
   Else if (N is an internal node with two children) {
       N will be replaced by the node N’ that has the
       next largest key after (or before) N.
   }

To find N’:
   1. Go to the right child of N
   2. Go down left until no left child is found.
   The node found is N’
Suppose we need to sort a list of employee records in ascending order, using the social security number (a 9-digit number) as the key (i.e., sort the records by social security number). If we need to guarantee that the running time will be no worse than $n \log n$, which sorting methods could we use?

- Merge sort
- Quicksort
- Insertion sort
- Either merge sort or quicksort
- None of these sorting algorithms
<table>
<thead>
<tr>
<th>Algo</th>
<th>Best case</th>
<th>Average case</th>
<th>Worst case</th>
</tr>
</thead>
<tbody>
<tr>
<td>MergeSort</td>
<td>$O( n \times \log(n) )$</td>
<td>$O( n \times \log(n) )$</td>
<td>$O( n \times \log(n) )$</td>
</tr>
<tr>
<td>QuickSort</td>
<td>$O( n \times \log(n) )$</td>
<td>$O( n \times \log(n) )$</td>
<td>$O( n^2 )$</td>
</tr>
<tr>
<td>InsertionSort</td>
<td>$O( n )$</td>
<td>$O( n^2 )$</td>
<td>$O( n^2 )$</td>
</tr>
<tr>
<td>HeapSort</td>
<td>$O( n \times \log(n) )$</td>
<td>$O( n \times \log(n) )$</td>
<td>$O( n \times \log(n) )$</td>
</tr>
<tr>
<td>BubbleSort</td>
<td>$O( n )$</td>
<td>$O( n^2 )$</td>
<td>$O( n^2 )$</td>
</tr>
</tbody>
</table>
Merge Sort

Divide

Merge

1 2 3 4

1 2 3 4
Which of the following assertions are true? (multiple choices)

- Heaps are binary search trees.
- Heaps are binary trees.
- Heaps can be used to implement priority queues.
- Heaps can be used to implement lists.

The bar chart shows the percentage of responses:
- 20% (24.1%) for Heaps are binary search trees.
- 60% (69.9%) for Heaps are binary trees.
- 71% (85.5%) for Heaps can be used to implement priority queues.
- 32% (38.6%) for Heaps can be used to implement lists.
A **heap** is a binary tree such that:

- For any node $n$ other than the root, $\text{key}(n) \geq \text{key}(\text{parent}(n))$

- Let $h$ be the height of the heap
  - First $h-1$ levels are full:
    - For $i = 0,...,h-1$, there are $2^i$ nodes of depth $i$
  - At depth $h$, the leaves are packed on the left side of the tree
What is the time-complexity of the removal of the highest priority key in a heap (where n is the number of keys stored)?

- \(O(1)\)
- \(O(n)\)
- \(O(\log(n))\)
Heaps: RemoveMin()

• The minimum key is always at the root of the heap!
• Replace the root with last node

• Restore heap-order property (see next)
Heaps: Bubbling-down

Restoring the heap-order property:

- Keep swapping the node with its smallest child as long as the node’s key is larger than it’s child’s key

Running time?

\[ O(h) = O(\log(n)) \]
You are using a hash table to store keys. Assuming there is **no collision**, which of the following operations have a $O(1)$ time-complexity? (multiple choices)

- Insert key.
- Remove key.
- Find key.
Which of the following assertions are true?* (multiple choices)

- Graphs are trees.
- Trees are graphs.
- A graph that is not a tree has at least one cycle.**
- An Hamiltonian cycle visits each vertex exactly once.
- An Eulerian cycle visits each vertex exactly once.

* The graph is connected, undirected and #nodes > 2.
** True only if edges are undirected.
In the graph shown above, starting from the green node at the top, which algorithm will visit the least number of nodes before visiting the yellow goal node?

• Depth First Search (DFS)
• **Breadth First Search (BFS)**
• BFS and DFS encounter same number of nodes before encounter the goal node.
Depth-First Search

Idea:

• Search "deeper" in the graph whenever possible
• Start at some vertex v
• After visiting vertex v, the next vertex to be explored is the first unvisited neighbor of v
• If v has no neighbor or if all its neighbors have explored, backtrack to the vertex from which we reached v
• Corresponds to adventurous web browsing: always click the first unvisited link available. Click "back" when you hit a dead-end.
Breadth-First Search

Idea:

- Explore graph layers by layers
- Start at some vertex $v$
- Then explore all the neighbors of $v$
- Then explore all the unvisited neighbors of the neighbors of $v$
- Then explore all the unvisited neighbors of the neighbors of the neighbors of $v$
- until no more unvisited vertices remain
T(n) = T(n-1) + O(n) is a recurrence for the running time of?

- **Insertion sort**
- Merge sort
- Quick sort
- Bubble sort

![Pie chart showing distribution of algorithms](image.png)

- Insertion sort: 18.3%
- Merge sort: 41.5%
- Quick sort: 26.8%
- Bubble sort: 13.4%
InsertionSort

for i ← 1 to length(A) − 1
    j ← i
    while j > 0 and A[j−1] > A[j]
        swap A[j] and A[j−1]
        j ← j − 1

Example: 3 7 4 9 5 2 6 1
          3 7 4 9 5 2 6 1
          3 7 4 9 5 2 6 1
          3 4 7 9 5 2 6 1
          3 4 7 9 5 2 6 1
          3 4 5 7 9 2 6 1
          2 3 4 5 7 9 6 1
          2 3 4 5 6 7 9 1
          1 2 3 4 5 6 7 9

The probability of team A winning any game is $\frac{1}{3}$. Team A plays team B in a tournament. If either team wins two games in a row, that team is declared the winner. At most three games are played in the tournament and, if no team has won the tournament at the end of three games, the tournament is declared a draw. What is the expected number of games in the tournament?

- $3$
- $\frac{19}{9}$
- $\boxed{\frac{22}{9}}$
- $\frac{25}{9}$
- $\frac{61}{27}$
What are the possible outcomes?

\[ E = l(AA) \cdot P(AA) + l(BAA) \cdot P(BAA) + l(BB) \cdot P(BB) + l(ABB) \cdot P(ABB) + l(ABA) \cdot P(ABA) + l(BAB) \cdot P(BAB) \]

\[ = \frac{2}{9} \cdot \frac{1}{9} + 3 \cdot \frac{2}{27} + 2 \cdot \frac{4}{9} + 3 \cdot \frac{4}{27} + 3 \cdot \frac{2}{27} + 3 \cdot \frac{4}{27} \]

\[ = \frac{22}{9} \]
Background

Expectation & Indicators
Expectation

- **Average or mean**

- The expected value of a discrete random variable $X$ is
  \[ E[X] = \sum_x x \Pr\{X=x\} \]

- **Linearity of Expectation**
  - $E[X+Y] = E[X] + E[Y]$, for all $X, Y$
  - $E[aX+Y] = a \cdot E[X] + E[Y]$, for constant $a$ and all $X, Y$

- For **mutually independent random variables** $X_1, \ldots, X_n$
  - $E[X_1X_2 \ldots X_n] = E[X_1] \cdot E[X_2] \cdot \ldots \cdot E[X_n]$
Expectation – Example

• Let $X$ be the RV denoting the value obtained when a fair die is thrown. What will be the mean of $X$, when the die is thrown $n$ times.
  
  – Let $X_1, X_2, \ldots, X_n$ denote the values obtained during the $n$ throws.
  
  – The mean of the values is $\frac{(X_1+X_2+\ldots+X_n)}{n}$.
  
  – Since the probability of getting values 1 thru 6 is $\frac{1}{6}$, on an average we can expect each of the 6 values to show up $\frac{1}{6}n$ times.
  
  – So, the numerator in the expression for mean can be written as $\frac{1}{6}n \cdot 1 + \frac{1}{6}n \cdot 2 + \ldots + \frac{1}{6}n \cdot 6$
  
  – The mean, hence, reduces to $\frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \ldots + \frac{1}{6} \cdot 6$, which is what we get if we apply the definition of expectation.
Indicator Random Variables

• A simple yet powerful technique for computing the expected value of a random variable.
• Convenient method for converting between probabilities and expectations.
• Helpful in situations in which there may be dependence.
• Takes only 2 values, 1 and 0.
• Indicator Random Variable for an event $A$ of a sample space is defined as:

$$I\{A\} = \begin{cases} 
1 & \text{if } A \text{ occurs,} \\
0 & \text{if } A \text{ does not occur.}
\end{cases}$$
Indicator Random Variable

**Lemma 5.1**

Given a sample space $S$ and an event $A$ in the sample space $S$, let $X_A = I\{A\}$. Then $E[X_A] = \Pr\{A\}$.

**Proof:**

Let $\bar{A} = S - A$ (Complement of $A$)

Then,

$$E[X_A] = E[I\{A\}]$$

$$= 1 \cdot \Pr\{A\} + 0 \cdot \Pr\{\bar{A}\}$$

$$= \Pr\{A\}$$
Problem: Determine the expected number of heads in $n$ coin flips.

Method 1: Without indicator random variables. Let $X$ be the random variable for the number of heads in $n$ flips. Then, $E[X] = \sum_{k=0..n} k \cdot \Pr\{X=k\}$

We can solve this with a lot of math.
Indicators RV – Example

- **Method 2**: Use Indicator Random Variables
- Define \( n \) indicator random variables, \( X_i, 1 \leq i \leq n \).
- Let \( X_i \) be the indicator random variable for the event that the \( i \)th flip results in a Head.
- \( X_i = I\{\text{the } i\text{th flip results in } H\} \)
- Then \( X = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^n X_i \).
- By Lemma 5.1, \( E[X_i] = \Pr\{H\} = \frac{1}{2}, 1 \leq i \leq n \).
- Expected number of heads is \( E[X] = E[\sum_{i=1}^n X_i] \).
- By linearity of expectation, \( E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i] \).
- \( E[X] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n \frac{1}{2} = n/2. \)
Supplement

Answers to java questions
You read the following statement in a Java program that compiles and executes:

    submarine.dive(depth);

- depth must be an int.
- **dive must be a method.**
- submarine must be the name of a class.
- submarine must be a method.
Consider the following program:
```java
public class MyClass{
    public MyClass() { /*code*/ };
    // more code...
}
```

How would you instantiate MyClass?

- `MyClass mc = new MyClass();`
- `MyClass mc = MyClass();`
- `MyClass mc = new MyClass;`
- `MyClass mc = new MyClass;`
- It can't be done. The constructor of MyClass should be defined as: `public void MyClass() { /*code*/ }`
You want to initialize all of the elements of a double array `a` to the same value equal to 1.5. What could you write? Assume that the array has been correctly initialized.

- `for(int i=1; i<a.length; i++) a[i] = 1.5;`
- `for(int i=0; i<=a.length; i++) a[i] = 1.5;`
- `for(int i=0; i<a.length; i++) a[i] = 1.5;`
- `for(int i=0; i<a.length+1; i++) a[i] = 1.5;`
- `for(int i=0; i<a.length-1; i++) a[i] = 1.5;`