CSE 250
Data Structures

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Lec 31: Expected Runtime

Warm-Up Question

What sorting algorithms have we seen, what are their complexities, and what benefits can we get from data that is sorted?

Warm-Up Question

What sorting algorithms have we seen, what are their complexities, and what benefits can we get from data that is sorted?

BubbleSort, SelectionSort, InsertionSort - $O(n^2)$

MergeSort, HeapSort - O(n log n)

Why Sort? Searching sorted data can be faster (binary vs linear search)

Announcements

- Midterm 2 Grading in Progress
- PA3 & WA5 coming soon!

Recap: Merge Sort

Divide: Split the sequence in half

$D(n) = \Theta(n)$ (can do in $\Theta(1)$)

Conquer: Sort the left and right halves

$a = 2, b = 2, c = 1$

Combine: Merge halves together

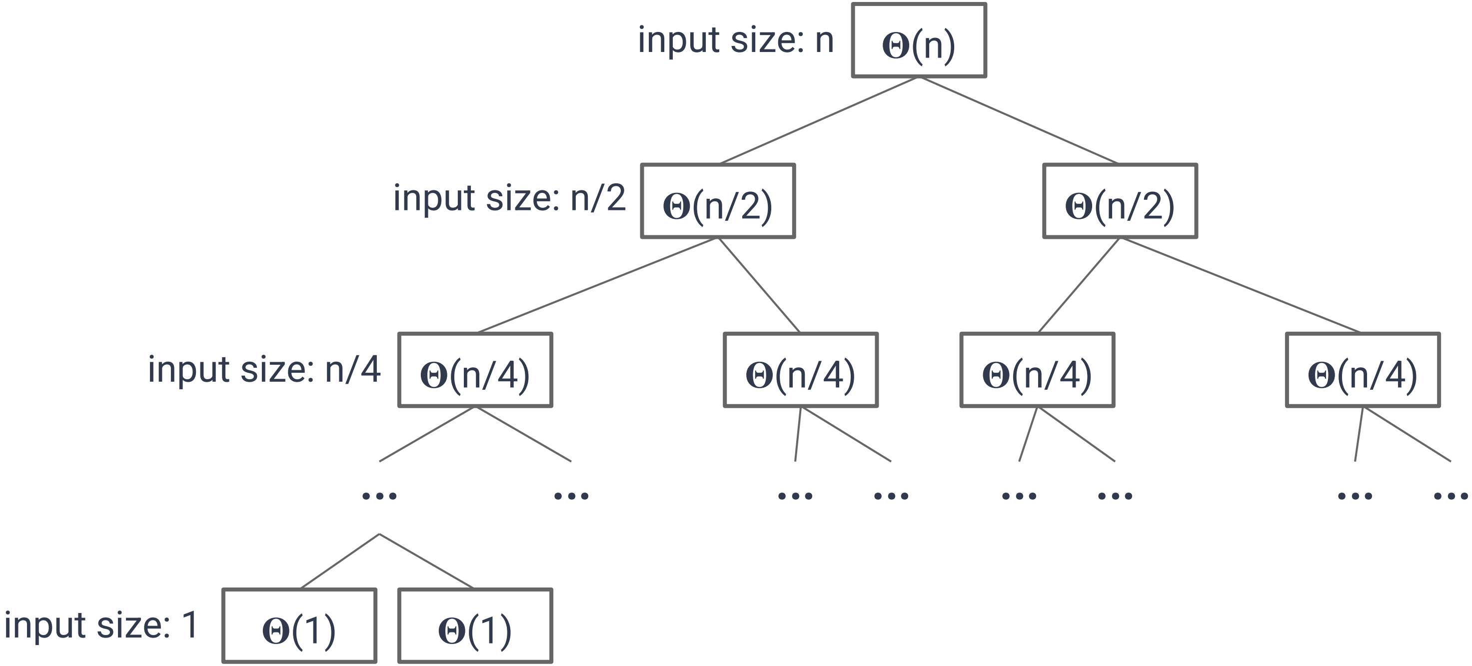
$C(n) = \Theta(n)$

Merge Sort: Growth Function

$$T(n) = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ 2 \cdot T(\frac{n}{2}) + \Theta(1) + \Theta(n) & \textbf{otherwise} \end{cases}$$

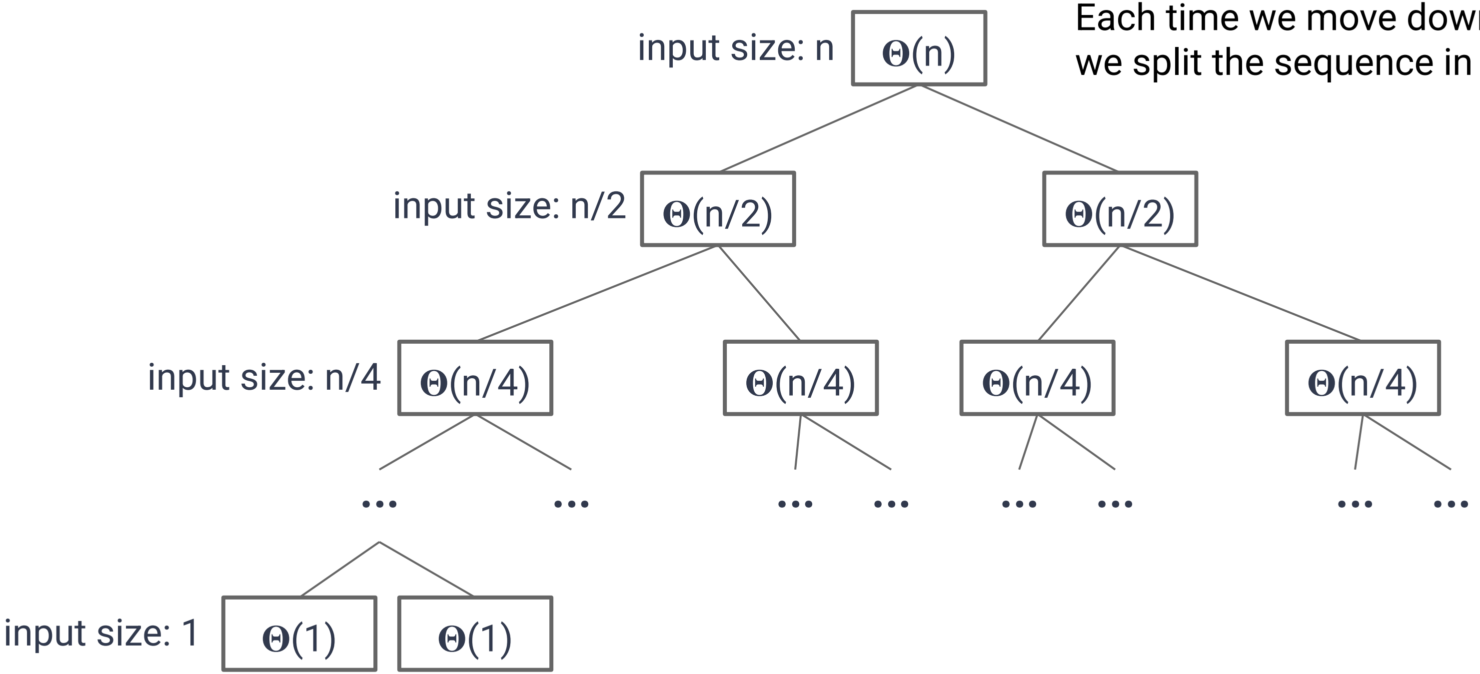
How do we find a closed-form hypothesis?

Merge Sort: Recursion Tree

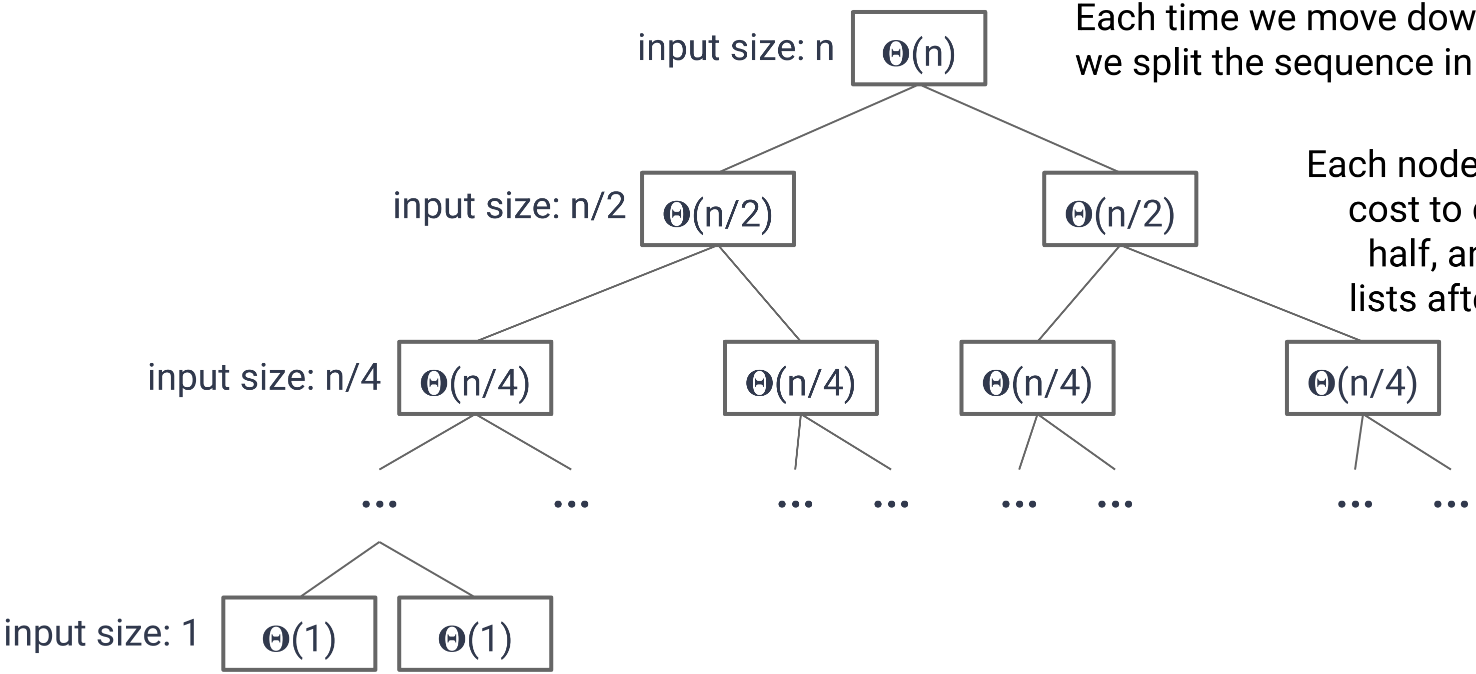


Merge Sort: Recursion Tree

Each time we move down a level, we split the sequence in half



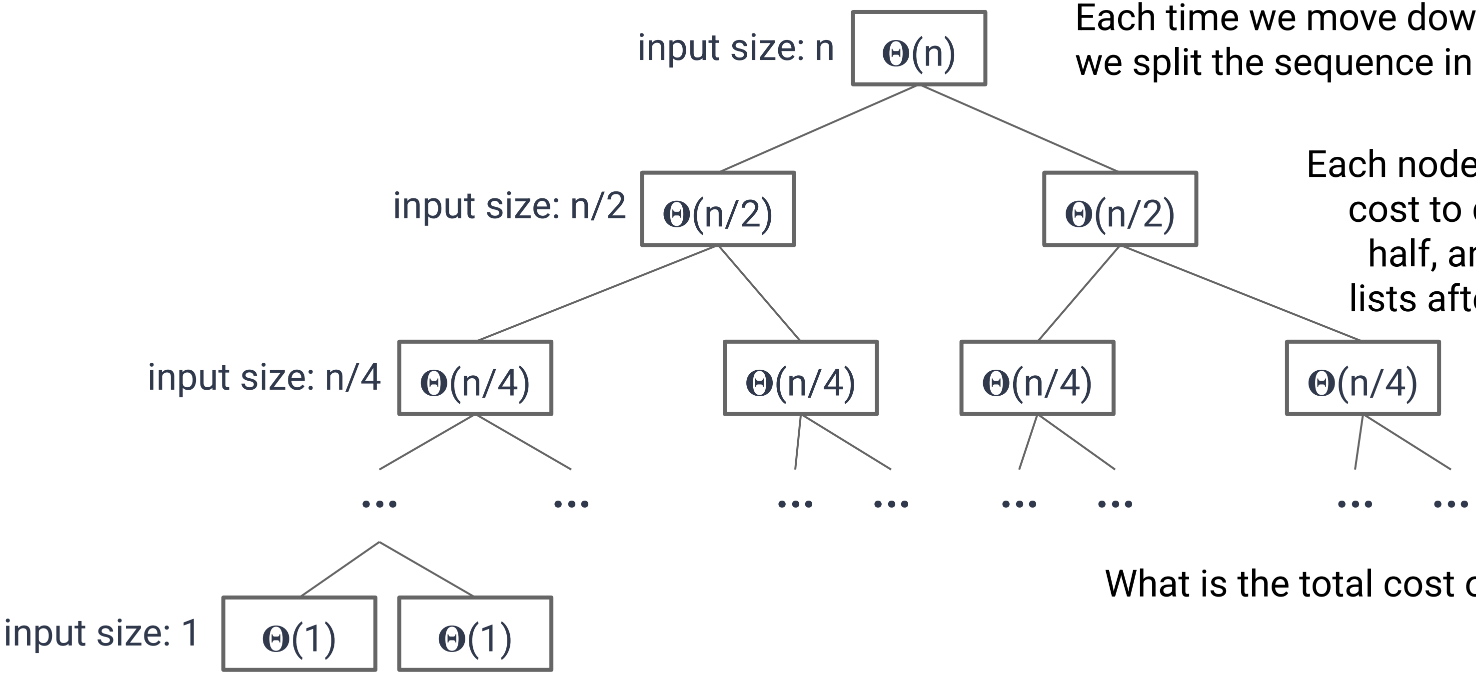
Merge Sort: Recursion Tree



Each time we move down a level, we split the sequence in half

Each node is labeled with the total cost to dividing the sequence in half, and combining the sorted lists after they are sorted by the lower levels

Merge Sort: Recursion Tree

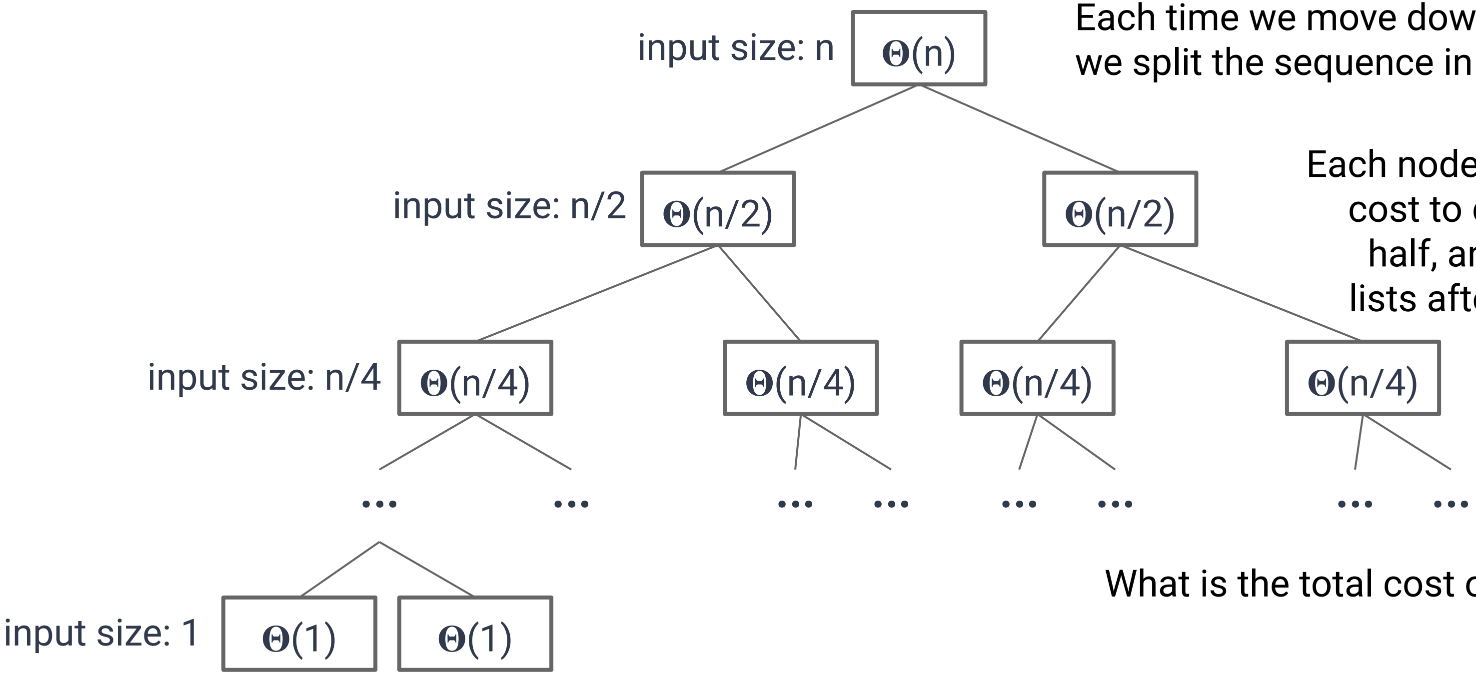


Each time we move down a level, we split the sequence in half

Each node is labeled with the total cost to dividing the sequence in half, and combining the sorted lists after they are sorted by the lower levels

What is the total cost of each level?

Merge Sort: Recursion Tree



Each time we move down a level, we split the sequence in half

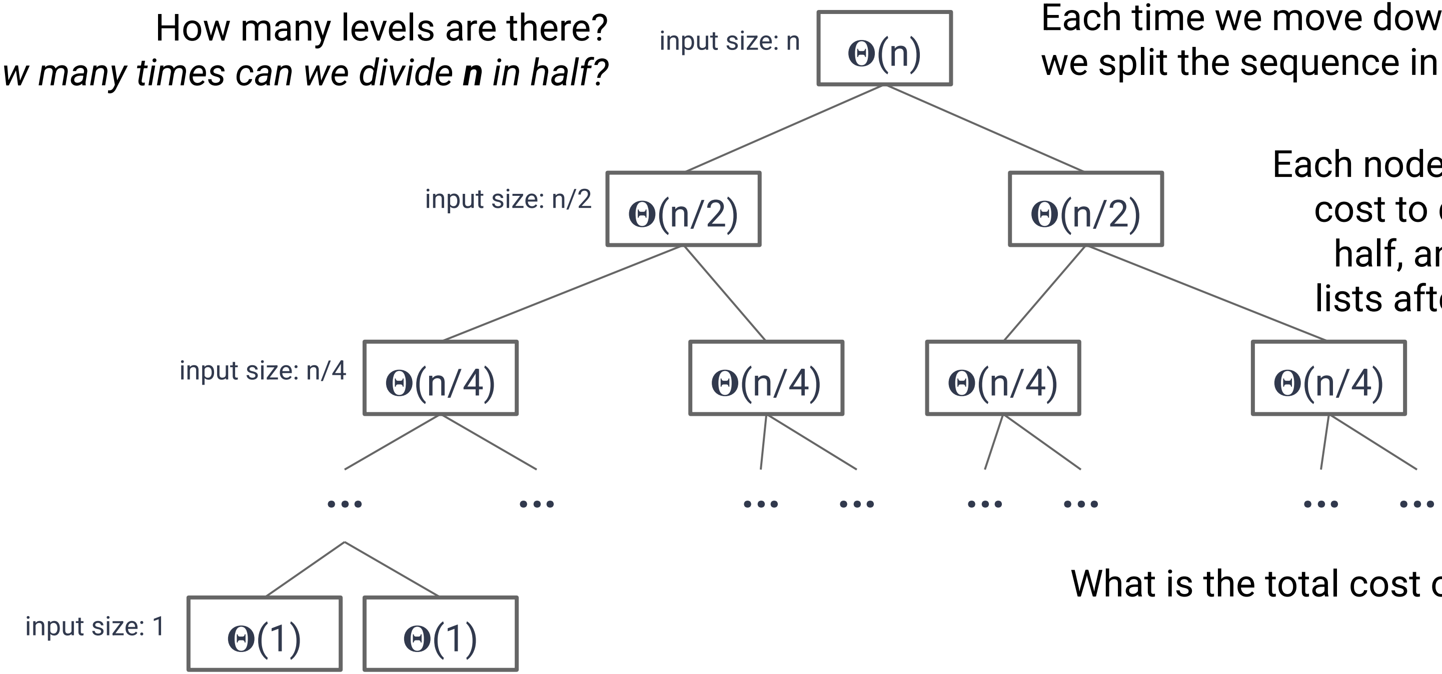
Each node is labeled with the total cost to dividing the sequence in half, and combining the sorted lists after they are sorted by the lower levels

What is the total cost of each level? $\Theta(n)$

Merge Sort: Recursion Tree

How many levels are there?
*How many times can we divide **n** in half?*

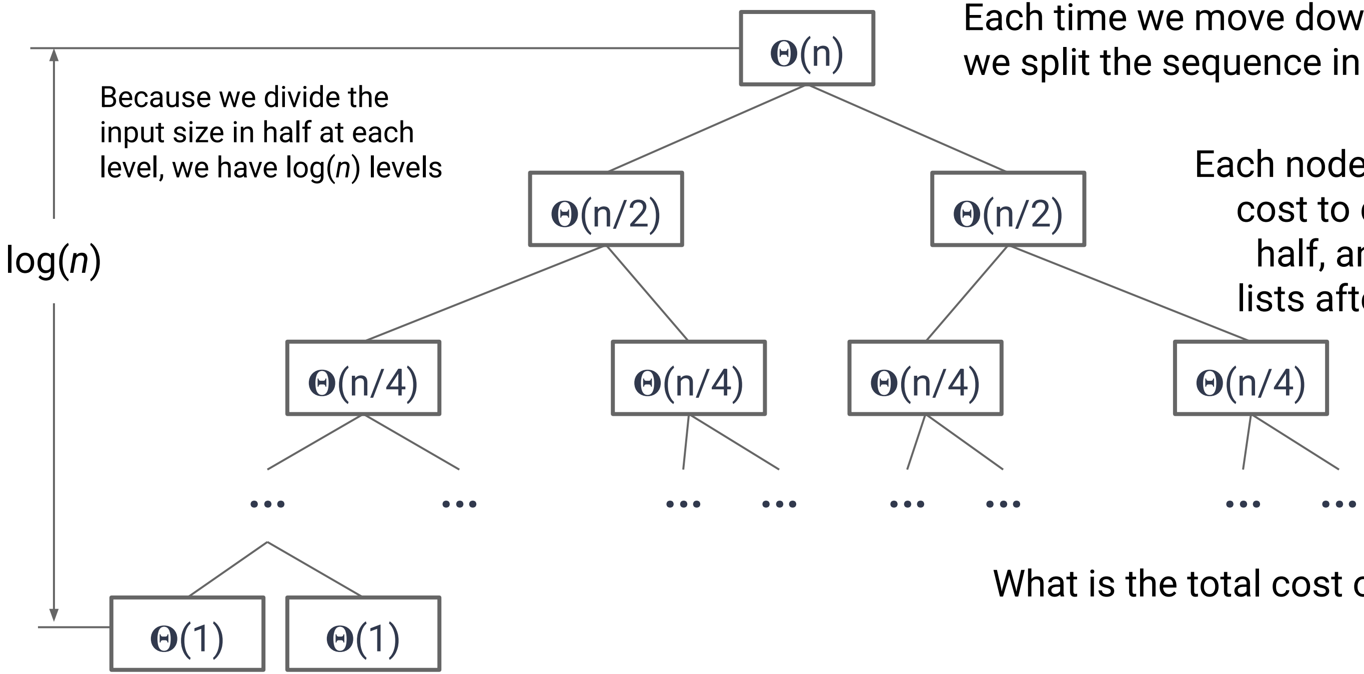
Each time we move down a level, we split the sequence in half



Each node is labeled with the total cost to dividing the sequence in half, and combining the sorted lists after they are sorted by the lower levels

What is the total cost of each level? $\Theta(n)$

Merge Sort: Recursion Tree



Each time we move down a level, we split the sequence in half

Each node is labeled with the total cost to dividing the sequence in half, and combining the sorted lists after they are sorted by the lower levels

What is the total cost of each level? $\Theta(n)$

Hypothesis: The cost of merge sort is $n \log(n)$

Merge Sort: Proof by Induction

Base Case: $T(1) \le c\ 1 \log(1)$

~~$c_0 \le 0$~~

$T(2) \le c\ 2 \log(2)$

$2c_0 + c_1 + 2c_2 \le 2c$

True when $c = c_0 + c_1 + c_2$

Merge Sort: Proof by Induction

Assume: $T(n/2) \leq c\,(n/2)\log(n/2)$

Show: $T(n) \leq cn\log(n)$

Merge Sort: Proof by Induction

Assume: $T(n/2) \leq c\ (n/2) \log(n/2)$

Show: $T(n) \leq cn \log(n)$

$$2 \cdot T(\frac{n}{2}) + c_1 + c_2 n \leq cn \log(n)$$

Merge Sort: Proof by Induction

Assume: $T(n/2) \leq c\ (n/2) \log(n/2)$

Show: $T(n) \leq cn \log(n)$

$$2 \cdot T(\frac{n}{2}) + c_1 + c_2 n \leq cn \log(n)$$

By the assumption, and transitivity, we just need to show:

$$2c\frac{n}{2} \log\left(\frac{n}{2}\right) + c_1 + c_2 n \leq cn \log(n)$$

Merge Sort: Proof by Induction

Assume: $T(n/2) \leq c\,(n/2)\log(n/2)$

Show: $T(n) \leq cn\log(n)$

$$2 \cdot T(\frac{n}{2}) + c_1 + c_2 n \leq cn\log(n)$$

By the assumption, and transitivity, we just need to show:

$$2c\frac{n}{2}\log\left(\frac{n}{2}\right) + c_1 + c_2 n \leq cn\log(n)$$

$$cn\log(n) - cn\log(2) + c_1 + c_2 n \leq cn\log(n)$$

Merge Sort: Proof by Induction

Assume: $T(n/2) \leq c\,(n/2)\log(n/2)$

Show: $T(n) \leq cn\log(n)$

$$2 \cdot T(\frac{n}{2}) + c_1 + c_2 n \leq cn\log(n)$$

By the assumption, and transitivity, we just need to show:

$$2c\frac{n}{2}\log\left(\frac{n}{2}\right) + c_1 + c_2 n \leq cn\log(n)$$

$$cn\log(n) - cn\log(2) + c_1 + c_2 n \leq cn\log(n)$$

$$c_1 + c_2 n \leq cn\log(2)$$

Merge Sort: Proof by Induction

$$c_1 + c_2 n \leq cn\log(2)$$

Merge Sort: Proof by Induction

$$c_1 + c_2 n \leq cn \log(2)$$

$$\frac{c_1}{n \log(2)} + \frac{c_2}{\log(2)} \leq c$$

Merge Sort: Proof by Induction

$$c_1 + c_2 n \leq cn\log(2)$$

$$\frac{c_1}{n\log(2)} + \frac{c_2}{\log(2)} \leq c$$

Which is true for any

$$n_0 \geq \frac{c_1}{\log(2)} \quad \text{and} \quad c > \frac{c_2}{\log(2)} + 1$$

Merge Sort: Follow Up

Where is all of the "work" being done?

Merge Sort: Follow Up

Where is all of the "work" being done?

The combine step

Merge Sort: Follow Up

Where is all of the "work" being done?

The combine step

Can we put the work in the divide step instead?

QuickSort: Intuition

Divide: Move *small* elements to the left and *big* elements to the right

How do we define what is *big* and what is *small*?

QuickSort: Intuition

Divide: Move *small* elements to the left and *big* elements to the right

How do we define what is *big* and what is *small*?

Pick a pivot value

QuickSort: Intuition

Divide: Move *small* elements to the left and *big* elements to the right

How do we define what is *big* and what is *small*?

Pick a pivot value

[smaller than pivot], pivot, [larger than pivot]

QuickSort: Intuition

Divide: Move *small* elements to the left and *big* elements to the right

How do we define what is *big* and what is *small*?

Pick a pivot value

[smaller than pivot], pivot, [larger than pivot]

How do we pick a pivot?

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

If we pick 8, the median value, we'll end up dividing our list in half during the divide step

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, 2, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, <u>2</u>, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, 2, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, 2, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, 2, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [14, 13, 9, 12, 11, 10, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, 2, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [14, 13, 9, 12, 11, 10, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, 2, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [11, 10, 9], 12, [14, 13, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, 2, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [11, <u>10</u>, 9], 12, [14, 13, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, 2, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [11, 10, 9], 12, [14, 13, 15]

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, [14, 13, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, 2, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [11, 10, 9], 12, [14, 13, 15]

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, [<u>14</u>, 13, 15]

QuickSort: Ideal Example

[4, 1, 8, 13, 12, 6, 2, 14, 7, 9, 3, 5, 11, 10, 15]

[4, 1, 7, 3, 6, 2, 5], 8, [14, 13, 9, 12, 11, 10, 15]

[1, 2, 3], 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, [6, 7, 5], 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [14, 13, 9, 12, 11, 10, 15]

1, 2, 3, 4, 5, 6, 7, 8, [11, 10, 9], 12, [14, 13, 15]

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, [14, 13, 15]

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

QuickSort: Ideal Example

If our pivot was the median value, then our list would be split in half by the divide step, resulting in the same structure as MergeSort…

…However, once we finish recursively dividing, we are done! No need for a combine step at all!

QuickSort: Idealized Algorithm

To sort an array of size *n*:

1. Pick a *pivot* value (median?)
2. Swap values until:
   a. elements at [1, $n/2$) are ≤ pivot
   b. elements at [$n/2$, n) are > pivot
3. Recursively sort the lower half
4. Recursively sort the upper half

Great! So...how do we find the median...?

Great! So...how do we find the median...?

Finding the median takes O(*n* log(*n*)) for an unsorted array :(

**** Actually...it can be done in O(n) but with prohibitively high constant factors*

QuickSort: Hypothetical

Imagine a world where we can obtain an ideal pivot in $O(1)$.

Now what is our growth function?

QuickSort: Hypothetical

Imagine a world where we can obtain an ideal pivot in *O*(1).

Now what is our growth function?

$$T_{quicksort}(n) = \begin{cases} \Theta(1) & \textbf{if } n = 1 \\ 2 \cdot T(\frac{n}{2}) + \Theta(n) + 0 & \textbf{otherwise} \end{cases}$$

Divide cost is O(n), Combine cost is 0

QuickSort: Hypothetical

Imagine a world where we can obtain an ideal pivot in *O*(1).

Now what is our growth function?

$$T_{quicksort}(n) = \begin{cases} \Theta(1) & \textbf{if } n = 1 \\ 2 \cdot T(\frac{n}{2}) + \Theta(n) + 0 & \textbf{otherwise} \end{cases}$$

Compare to Merge Sort:

$$T_{mergesort}(n) = \begin{cases} \Theta(1) & \textbf{if } n = 1 \\ 2 \cdot T(\frac{n}{2}) + \Theta(1) + \Theta(n) & \textbf{otherwise} \end{cases}$$

QuickSort: Attempt #2

So how can we pick a pivot value (in $O(1)$ time)? (even if it's not ideal)

QuickSort: Attempt #2

So how can we pick a pivot value (in O(1) time)? (even if it's not ideal)

Idea: Pick it randomly! On average, half the values will be lower.

QuickSort: Attempt #2

To sort an array of size *n*:

1. Pick a value at random as the *pivot*
2. Swap values until the array is subdivided into:
    a. *low*: array elements < *pivot*
    b. *pivot*
    c. *high:* array elements > pivot
3. Recursively sort *low*
4. Recursively sort *high*

QuickSort: Runtime

What is the worst-case runtime?

QuickSort: Worst-Case Scenario

What if we always pick the worst pivot?

`[8,7,6,5,4,3,2,1]`

QuickSort: Worst-Case Scenario

What if we always pick the worst pivot?

`[8,7,6,5,4,3,2,1]`

`[7,6,5,4,3,2,1],8,[]`

QuickSort: Worst-Case Scenario

What if we always pick the worst pivot?

```
[8,7,6,5,4,3,2,1]

[7,6,5,4,3,2,1],8,[]

[6,5,4,3,2,1],7,[],8
```

QuickSort: Worst-Case Scenario

What if we always pick the worst pivot?

```
[8,7,6,5,4,3,2,1]

[7,6,5,4,3,2,1],8,[]

[6,5,4,3,2,1],7,[],8

[5,4,3,2,1],6,[],7,8
```

QuickSort: Worst-Case Scenario

What if we always pick the worst pivot?

```
[8,7,6,5,4,3,2,1]

[7,6,5,4,3,2,1],8,[]

[6,5,4,3,2,1],7,[],8

[5,4,3,2,1],6,[],7,8

...
```

QuickSort: Worst-Case Runtime

What is the worst-case runtime?

QuickSort: Worst-Case Runtime

What is the worst-case runtime?

$$T_{quicksort}(n) \in O(n^2)$$

QuickSort: Worst-Case Runtime

What is the worst-case runtime?

$$T_{quicksort}(n) \in O(n^2)$$

Remember: This is called the unqualified runtime…we don't take any extra context into account

QuickSort: Worst-Case Runtime

Is the worst case runtime representative?

QuickSort: Worst-Case Runtime

Is the worst case runtime representative?

No! (the actual runtime will almost always be faster)

QuickSort: Worst-Case Runtime

Is the worst case runtime representative?

No! (the actual runtime will almost always be faster)

But what **can** we say about runtime?

QuickSort

Let's say we pick Xth largest element for our pivot.

What is the runtime ($T(n)$)?

QuickSort

Let's say we pick Xth largest element for our pivot.

What is the runtime ($T(n)$)?

There are ***n*** possible outcomes, ranging from picking the ideal (median) to the worst case (biggest or smallest)

$$\begin{cases} T(0) + T(n-1) + \Theta(n) & \textbf{if } X = 1 \\ T(1) + T(n-2) + \Theta(n) & \textbf{if } X = 2 \\ T(2) + T(n-3) + \Theta(n) & \textbf{if } X = 3 \\ .. & \\ T(n-2) + T(1) + \Theta(n) & \textbf{if } X = n-1 \\ T(n-1) + T(0) + \Theta(n) & \textbf{if } X = n \end{cases}$$

Probabilities

How likely are we to pick $X = k$ for any specific k?

Probability Theory (Great Class...)

If I roll a d6 (6-sided die) *x* times,

what is the average roll over all possible outcomes?

A single die roll

If I rolled a d6 1 time…

Roll	Probability	Outcome
⚀	1/6	1
⚁	1/6	2
⚂	1/6	3
⚃	1/6	4
⚄	1/6	5
⚅	1/6	6

Expected Value

The **Expected Value** of a random variable (ie the number rolled on the d6) is the sum of all outcomes times the probability of that outcome

$$\sum_i Probability_i \cdot Contribution_i$$

Expected Value

The **Expected Value** of a random variable (ie the number rolled on the d6) is the sum of all outcomes times the probability of that outcome

$$\sum_{i=1}^{6} \frac{1}{6} i = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$$

Expected Value

The **Expected Value** of a random variable (ie the number rolled on the d6) is the sum of all outcomes times the probability of that outcome

$$\sum_{i=1}^{6} \frac{1}{6} i = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$$

We refer to the expected value of a random variable as ***E[X]***

Expected Value

If I roll a 6-sided die, the probability of a particular side being rolled is ⅙

If *X* is a random variable representing this die roll, then the expected value of *X* is:

$$E[X] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6$$

$$E[X] = \sum_{i=1}^{6} \frac{1}{6} i = 3.5$$

Expected Value

If I roll a 20-sided die, the probability of a particular side being rolled is 1/20

If *X* is a random variable representing this die roll, then the expected value of *X* is:

$$E[X] = \frac{1}{20} \cdot 1 + \frac{1}{20} \cdot 2 + \ldots + \frac{1}{20} \cdot 20 = \sum_{i=1}^{20} \frac{1}{20} i$$

Expected Value

If I roll an *n*-sided die, the probability of a particular side being rolled is 1/*n*

If *X* is a random variable representing this die roll, then the expected value of *X* is:

$$E[X] = \frac{1}{n} \cdot 1 + \frac{1}{n} \cdot 2 + ... + \frac{1}{n} \cdot n = \sum_{i=1}^{n} \frac{1}{n} i$$

$$E[X] = \sum_{i} P_i \cdot X_i$$

Linearity of Expectation

Expected Value is Linear; ie:

$E[X+Y] = E[X] + E[Y]$ and $E[cX] = cE[X]$

Linearity of Expectation

Expected Value is Linear; ie:

$E[X+Y] = E[X] + E[Y]$ and $E[cX] = cE[X]$

What if we roll a d6 twice? What do we expect the sum to be?

Linearity of Expectation

Expected Value is Linear; ie:

$E[X+Y] = E[X] + E[Y]$ and $E[cX] = cE[X]$

What if we roll a d6 twice? What do we expect the sum to be?

If ***X*** and ***Y*** are our dice rolls, $E[X + Y] = E[X] + E[Y] = 3.5 + 3.5 = 7$

or alternatively

$E[2X] = 2E[X] = 2 * 3.5 = 7$

Probabilities

How likely are we to pick $X = k$ for any specific k?

Probabilities

How likely are we to pick X = k for any specific k?

$P[X = k] = 1/n$

…Picking a pivot is like rolling an n-sided die

QuickSort Runtime

Now we can write our runtime function in terms of random variables:

$$T(n) = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ T(0) + T(n-1) + \Theta(n) & \textbf{if } n > 1 \land X = 1 \\ T(1) + T(n-2) + \Theta(n) & \textbf{if } n > 1 \land X = 2 \\ T(2) + T(n-3) + \Theta(n) & \textbf{if } n > 1 \land X = 3 \\ .. & \\ T(n-2) + T(1) + \Theta(n) & \textbf{if } n > 1 \land X = n-1 \\ T(n-1) + T(0) + \Theta(n) & \textbf{if } n > 1 \land X = n \end{cases}$$

QuickSort Runtime

Now we can write our runtime function in terms of random variables:

This would be our runtime if we randomly pick the smallest pivot

$$T(n) = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ \boxed{T(0) + T(n-1) + \Theta(n)} & \textbf{if } n > 1 \land X = 1 \\ T(1) + T(n-2) + \Theta(n) & \textbf{if } n > 1 \land X = 2 \\ T(2) + T(n-3) + \Theta(n) & \textbf{if } n > 1 \land X = 3 \\ .. & \\ T(n-2) + T(1) + \Theta(n) & \textbf{if } n > 1 \land X = n-1 \\ T(n-1) + T(0) + \Theta(n) & \textbf{if } n > 1 \land X = n \end{cases}$$

QuickSort Runtime

Now we can write our runtime function in terms of random variables:

This would be our runtime if we randomly pick the second smallest pivot

$$T(n) = \begin{cases} \Theta(1) & \text{if } n \leq 1 \\ T(0) + T(n-1) + \Theta(n) & \text{if } n > 1 \wedge X = 1 \\ \boxed{T(1) + T(n-2) + \Theta(n)} & \text{if } n > 1 \wedge X = 2 \\ T(2) + T(n-3) + \Theta(n) & \text{if } n > 1 \wedge X = 3 \\ .. & \\ T(n-2) + T(1) + \Theta(n) & \text{if } n > 1 \wedge X = n-1 \\ T(n-1) + T(0) + \Theta(n) & \text{if } n > 1 \wedge X = n \end{cases}$$

QuickSort Runtime

Now we can write our runtime function in terms of random variables:

This would be our runtime if we randomly pick the third smallest pivot

$$T(n) = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ T(0) + T(n-1) + \Theta(n) & \textbf{if } n > 1 \wedge X = 1 \\ T(1) + T(n-2) + \Theta(n) & \textbf{if } n > 1 \wedge X = 2 \\ \boxed{T(2) + T(n-3) + \Theta(n)} & \textbf{if } n > 1 \wedge X = 3 \\ .. & \\ T(n-2) + T(1) + \Theta(n) & \textbf{if } n > 1 \wedge X = n-1 \\ T(n-1) + T(0) + \Theta(n) & \textbf{if } n > 1 \wedge X = n \end{cases}$$

QuickSort Runtime

Now we can write our runtime function in terms of random variables:

This would be our runtime if we randomly pick the third smallest pivot

...etc

and each pivot has a 1/n chance of being selected

$$T(n) = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ T(0) + T(n-1) + \Theta(n) & \textbf{if } n > 1 \wedge X = 1 \\ T(1) + T(n-2) + \Theta(n) & \textbf{if } n > 1 \wedge X = 2 \\ T(2) + T(n-3) + \Theta(n) & \textbf{if } n > 1 \wedge X = 3 \\ .. & \\ T(n-2) + T(1) + \Theta(n) & \textbf{if } n > 1 \wedge X = n-1 \\ T(n-1) + T(0) + \Theta(n) & \textbf{if } n > 1 \wedge X = n \end{cases}$$

QuickSort Runtime

…and convert it to the **<u>expected runtime</u>** over the variable ***X***

$$E[T(n)] = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ E[T(X-1) + T(n-X)] + \Theta(n) & \textbf{otherwise} \end{cases}$$

QuickSort Runtime

…and convert it to the **expected runtime** over the variable ***X***

$$E[T(n)] = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ E[T(X-1) + T(n-X)] + \Theta(n) & \textbf{otherwise} \end{cases}$$

This growth function represents the **expected** number of steps we must take to sort using QuickSort…and just like any other growth function, we can find O, **Ω**, and potentially **Θ** bounds

QuickSort Runtime

$$E[T(n)] = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ E[T(X-1) + T(n-X)] + \Theta(n) & \textbf{otherwise} \end{cases}$$

Expected value is linear, so we can be split up

QuickSort Runtime

$$E[T(n)] = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ E[T(X-1)] + E[T(n-X)] + \Theta(n) & \textbf{otherwise} \end{cases}$$

QuickSort Runtime

$$E[T(n)] = \begin{cases} \Theta(1) & \textbf{if } n \le 1 \\ E[T(X-1)] + E[T(n-X)] + \Theta(n) & \textbf{otherwise} \end{cases}$$

How are these two terms related?

QuickSort Runtime

$$E[T(X-1)]$$

QuickSort Runtime

$$E[T(X-1)]$$

$$= \sum_{i=1}^{n} P_i \cdot T(X_i - 1)$$

QuickSort Runtime

$$E[T(X-1)]$$

$$= \sum_{i=1}^{n} P_i \cdot T(X_i - 1)$$

$$= \sum_{i=1}^{n} \frac{1}{n} \cdot T(i-1)$$

QuickSort Runtime

$$E[T(X-1)]$$

$$= \sum_{i=1}^{n} P_i \cdot T(X_i - 1)$$

$$= \sum_{i=1}^{n} \frac{1}{n} \cdot T(i - 1)$$

$$= \sum_{i=1}^{n} \frac{1}{n} \cdot T(n - i)$$

QuickSort Runtime

$$E[T(X-1)]$$

$$= \sum_{i=1}^{n} P_i \cdot T(X_i - 1)$$

$$= \sum_{i=1}^{n} \frac{1}{n} \cdot T(i-1)$$

$$= \sum_{i=1}^{n} \frac{1}{n} \cdot T(n-i) = E[T(n-X)]$$

QuickSort Runtime

$$E[T(X-1)]$$

$$= \sum_{i=1}^{n} P_i \cdot T(X_i - 1)$$

$$= \sum_{i=1}^{n} \frac{1}{n} \cdot T(i-1)$$

$$= \sum_{i=1}^{n} \frac{1}{n} \cdot T(n-i) = E[T(n-X)]$$

They are equivalent!!

QuickSort Runtime

$$E[T(n)] = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ 2E[T(X-1)] + \Theta(n) & \textbf{otherwise} \end{cases}$$

QuickSort Runtime

$$E[T(n)] = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ \boxed{2E[T(X-1)]} + \Theta(n) & \textbf{otherwise} \end{cases}$$

This is a summation of multiple random variables, and expectation is linear

QuickSort Runtime

$$E[T(n)] = \begin{cases} \Theta(1) & \textbf{if } n \leq 1 \\ \frac{2}{n}\left(\sum_{i=0}^{n-1} E[T(i)]\right) + \Theta(n) & \textbf{otherwise} \end{cases}$$

Back to Induction

Hypothesis: $E[T(n)] \in O(n \log(n))$

Note that our hypothesis is now about the EXPECTED runtime…that is what we are trying to prove

Base Case

Base Case: $E[T(2)] \leq c\ (2 \log(2))$

Base Case

Base Case: $E[T(2)] \leq c\,(2 \log(2))$

$2 \cdot E_i[T(i - 1)] + 2c_1 \leq 2c$

Base Case

Base Case: $E[T(2)] \leq c\ (2 \log(2))$

$$2 \cdot E_i[T(i - 1)] + 2c_1 \leq 2c$$

$$2 \cdot (T(0)/2 + T(1)/2) + 2c_1 \leq 2c$$

Base Case

Base Case: $E[T(2)] \leq c\ (2 \log(2))$

$$2 \cdot E_i[T(i - 1)] + 2c_1 \leq 2c$$

$$2 \cdot (T(0)/2 + T(1)/2) + 2c_1 \leq 2c$$

$$T(0) + T(1) + 2c_1 \leq 2c$$

Base Case

Base Case: $E[T(2)] \leq c\ (2 \log(2))$

$$2 \cdot E_i[T(i - 1)] + 2c_1 \leq 2c$$

$$2 \cdot (T(0)/2 + T(1)/2) + 2c_1 \leq 2c$$

$$T(0) + T(1) + 2c_1 \leq 2c$$

$$2c_0 + 2c_1 \leq 2c$$

Base Case

Base Case: $E[T(2)] \leq c\ (2 \log(2))$

$$2 \cdot E_i[T(i - 1)] + 2c_1 \leq 2c$$

$$2 \cdot (T(0)/2 + T(1)/2) + 2c_1 \leq 2c$$

$$T(0) + T(1) + 2c_1 \leq 2c$$

$$2c_0 + 2c_1 \leq 2c$$

True for any $c \geq c_0 + c_1$

Inductive Case

Assume: $E[T(n')] \le c\,(n' \log(n'))$ for **all** $n' < n$

Show: $E[T(n)] \le c\,(n \log(n))$

Inductive Case

Assume: $E[T(n')] \leq c\ (n' \log(n'))$ for **all** $n' < n$

Show: $E[T(n)] \leq c\ (n \log(n))$

$$\frac{2}{n}\left(\sum_{i=0}^{n-1} E[T(i)]\right) + c_1 \leq cn\log(n)$$

Inductive Case

Assume: $E[T(n')] \leq c\,(n' \log(n'))$ for **all** $n' < n$

Show: $E[T(n)] \leq c\,(n \log(n))$

Our *i* here is always less than *n*, so we can use our assumption to substitute

$$\frac{2}{n}\left(\sum_{i=0}^{n-1} E[T(i)]\right) + c_1 \leq cn\log(n)$$

$$\frac{2}{n}\left(\sum_{i=0}^{n-1} ci\log(i)\right) + c_1 \leq cn\log(n)$$

Inductive Case

Assume: $E[T(n')] \leq c\,(n' \log(n'))$ for **all** $n' < n$

Show: $E[T(n)] \leq c\,(n \log(n))$

$$\frac{2}{n}\left(\sum_{i=0}^{n-1} E[T(i)]\right) + c_1 \leq cn\log(n)$$

$$\frac{2}{n}\left(\sum_{i=0}^{n-1} ci\log(i)\right) + c_1 \leq cn\log(n)$$

$$c\frac{2}{n}\left(\sum_{i=0}^{n-1} i\log(n)\right) + c_1 \leq cn\log(n)$$

Inductive Case

$$c\frac{2}{n}\left(\sum_{i=0}^{n-1} i\log(n)\right) + c_1 \le cn\log(n)$$

Inductive Case

$$c\frac{2}{n}\left(\sum_{i=0}^{n-1} i\log(n)\right) + c_1 \leq cn\log(n)$$

$$c\frac{2\log(n)}{n}\left(\sum_{i=0}^{n-1} i\right) + c_1 \leq cn\log(n)$$

Inductive Case

$$c\frac{2}{n}\left(\sum_{i=0}^{n-1} i\log(n)\right) + c_1 \leq cn\log(n)$$

$$c\frac{2\log(n)}{n}\left(\sum_{i=0}^{n-1} i\right) + c_1 \leq cn\log(n)$$

$$c\frac{2\log(n)}{n}\left(\frac{(n-1)(n-1+1)}{2}\right) + c_1 \leq cn\log(n)$$

Inductive Case

$$c\frac{2}{n}\left(\sum_{i=0}^{n-1} i\log(n)\right) + c_1 \leq cn\log(n)$$

$$c\frac{2\log(n)}{n}\left(\sum_{i=0}^{n-1} i\right) + c_1 \leq cn\log(n)$$

$$c\frac{2\log(n)}{n}\left(\frac{(n-1)(n-1+1)}{2}\right) + c_1 \leq cn\log(n)$$

$$c\frac{\log(n)}{n}\left(n^2 - n\right) + c_1 \leq cn\log(n)$$

Inductive Case

$$c\frac{2}{n}\left(\sum_{i=0}^{n-1} i\log(n)\right) + c_1 \leq cn\log(n)$$

$$c\frac{2\log(n)}{n}\left(\sum_{i=0}^{n-1} i\right) + c_1 \leq cn\log(n)$$

$$c\frac{2\log(n)}{n}\left(\frac{(n-1)(n-1+1)}{2}\right) + c_1 \leq cn\log(n)$$

$$c\frac{\log(n)}{n}\left(n^2 - n\right) + c_1 \leq cn\log(n)$$

$$cn\log(n) - c\log(n) + c_1 \leq cn\log(n)$$

Inductive Case

$$c\frac{2}{n}\left(\sum_{i=0}^{n-1} i\log(n)\right) + c_1 \leq cn\log(n)$$

$$c\frac{2\log(n)}{n}\left(\sum_{i=0}^{n-1} i\right) + c_1 \leq cn\log(n)$$

$$c\frac{2\log(n)}{n}\left(\frac{(n-1)(n-1+1)}{2}\right) + c_1 \leq cn\log(n)$$

$$c\frac{\log(n)}{n}\left(n^2 - n\right) + c_1 \leq cn\log(n)$$

$$cn\log(n) - c\log(n) + c_1 \leq cn\log(n)$$

$$c_1 \leq c\log(n)$$

QuickSort

So…is QuickSort $O(n \log(n))$...?

No! It is <u>expected</u> to be, but that is not a guarantee

What guarantees do you get?

If *f*(*n*) is a Tight Bound
The algorithm always runs in $cf(n)$ steps

If *f*(*n*) is a Worst-Case Bound
The algorithm always runs in at most $cf(n)$

If *f*(*n*) is an Amortized Worst-Case Bound
n invocations of the algorithm **always** run in $cnf(n)$ steps

If *f*(*n*) is an Average/Expected Bound
…we don't have any guarantees

What guarantees do you get?

If $f(n)$ is a Tight Bound

The algorithm always runs in $cf(n)$ steps

← Unqualified runtime

If $f(n)$ is a Worst-Case Bound

The algorithm always runs in at most $cf(n)$

If $f(n)$ is an Amortized Worst-Case Bound

n invocations of the algorithm **always** run in $cnf(n)$ steps

If $f(n)$ is an Average/Expected Bound

…we don't have any guarantees