Lecture 9:
Linear Sorting

Steven Skiena

Department of Computer Science
State University of New York
Stony Brook, NY 11794–4400

http://www.cs.stonybrook.edu/~skiena

Problem of the Day

The *nuts and bolts* problem is defined as follows. You are given a collection of n bolts of different widths, and n corresponding nuts. You can test whether a given nut and bolt together, from which you learn whether the nut is too large, too small, or an exact match for the bolt. The differences in size between pairs of nuts or bolts can be too small to see by eye, so you cannot rely on comparing the sizes of two nuts or two bolts directly. You are to match each bolt to each nut.

1. Give an $O(n^2)$ algorithm to solve the nuts and bolts problem.
2. Suppose that instead of matching all of the nuts and bolts, you wish to find the smallest bolt and its corresponding nut. Show that this can be done in only $2n - 2$ comparisons.
3. Match the nuts and bolts in expected $O(n \log n)$ time.

Randomized Quicksort

Suppose you are writing a sorting program, to run on data given to you by your worst enemy. Quicksort is good on average, but bad on certain worst-case instances.

If you used Quicksort, what kind of data would your enemy give you to run it on? Exactly the worst-case instance, to make you look bad.

But suppose you picked the pivot element at *random*.

Now your enemy cannot design a worst-case instance to give to you, because no matter which data they give you, you would have the same probability of picking a good pivot!

Randomized Guarantees

Randomization is a very important and useful idea. By either picking a random pivot or scrambling the permutation before sorting it, we can say:

> "With high probability, randomized quicksort runs in $\Theta(n \lg n)$ time."

Where before, all we could say is:

> "If you give me random input data, quicksort runs in expected $\Theta(n \lg n)$ time."

Importance of Randomization

Since the time bound how does not depend upon your input distribution, this means that unless we are *extremely* unlucky (as opposed to ill prepared or unpopular) we will certainly get good performance.

Randomization is a general tool to improve algorithms with bad worst-case but good average-case complexity.

The worst-case is still there, but we almost certainly won't see it.

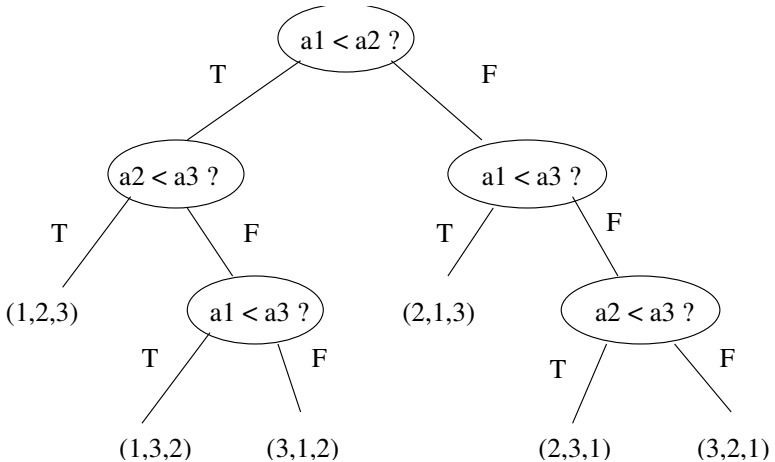
Is Quicksort really faster than Heapsort?

Since Heapsort is $\Theta(n \lg n)$ and selection sort is $\Theta(n^2)$, there is no debate about which will be better for decent-sized files.
When Quicksort is implemented well, it is typically 2-3 times faster than mergesort or heapsort.
The primary reason is that the operations in the innermost loop are simpler.
Since the difference between the two programs will be limited to a multiplicative constant factor, the details of how you program each algorithm will make a big difference.

Can we sort in $o(n \lg n)$?

Any comparison-based sorting program can be thought of as defining a decision tree of possible executions.

Running the same program twice on the same permutation causes it to do exactly the same thing, but running it on different permutations of the same data causes a different sequence of comparisons to be made on each.



Claim: the height of this decision tree is the worst-case complexity of sorting.

Lower Bound Analysis

Since any two different permutations of n elements requires a different sequence of steps to sort, there must be at least $n!$ different paths from the root to leaves in the decision tree.

Thus there must be at least $n!$ different leaves in this binary tree.

Since a binary tree of height h has at most 2^h leaves, we know $n! \leq 2^h$, or $h \geq \lg(n!)$.

By inspection $n! > (n/2)^{n/2}$, since the last $n/2$ terms of the product are each greater than $n/2$. Thus

$$\log(n!) > \log((n/2)^{n/2}) = n/2 \log(n/2) \rightarrow \Theta(n \log n)$$

Stirling's Approximation

By Stirling's approximation, a better bound is $n! > (n/e)^n$ where $e = 2.718$.

$$h \geq \lg(n/e)^n = n \lg n - n \lg e = \Omega(n \lg n)$$

Non-Comparison-Based Sorting

All the sorting algorithms we have seen assume binary comparisons as the basic primative, questions of the form "is x before y?".

But how would you sort a deck of playing cards?

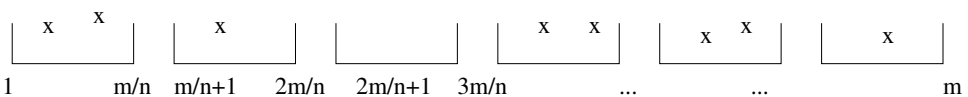
Most likely you would set up 13 piles and put all cards with the same number in one pile.

With only a constant number of cards left in each pile, you can use insertion sort to order by suite and concatenate everything together.

If we could find the correct pile for each card in constant time, and each pile gets $O(1)$ cards, this algorithm takes $O(n)$ time.

Bucketsort

Suppose we are sorting n numbers from 1 to m, where we know the numbers are approximately uniformly distributed. We can set up n buckets, each responsible for an interval of m/n numbers from 1 to m



Given an input number x, it belongs in bucket number $\lceil xn/m \rceil$.

If we use an array of buckets, each item gets mapped to the right bucket in $O(1)$ time.

Bucketsort Analysis

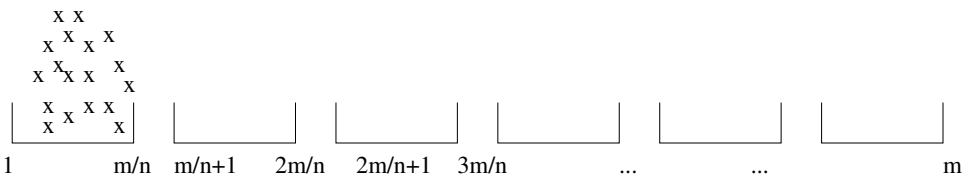
With uniformly distributed keys, the expected number of items per bucket is 1. Thus sorting each bucket takes $O(1)$ time!

The total effort of bucketing, sorting buckets, and concatenating the sorted buckets together is $O(n)$.

What happened to our $\Omega(n \lg n)$ lower bound!

Worst-Case vs. Assumed-Case

Bad things happen to bucketsort when we assume the wrong distribution.



We might spend linear time distributing our items into buckets and learn *nothing*.

Problems like this are why we worry about the worst-case performance of algorithms!

Real World Distributions

The worst case "shouldn't" happen if we understand the distribution of our data.
Consider the distribution of names in a telephone book.

- Will there be a lot of Skiena's?
- Will there be a lot of Smith's?
- Will there be a lot of Shifflett's?

Either make *sure* you understand your data, or use a good worst-case or randomized algorithm!

The Shifflett's of Charlottesville

For comparison, note that there are seven Shifflett's (of various spellings) in the 1000 page Manhattan telephone directory.

Shifflett DeDDie K Ruckersville 985-7957
Shifflett Debra S SR 617 Quinque 985-8813
Shifflett Delma SR609 985-3688
Shifflett Delmas Crozet 823-5901
Shifflett Dempsey & Marilynn
100 Greenbrier Ter 973-7195
Shifflett Denise Rt 627 Dyke 985-8097
Shifflett Dennis Stanardsville 985-4560
Shifflett Dennis H Stanardsville 985-2924
Shifflett Dewey E Rt667 985-6576
Shifflett Dewey O Dyke 985-7249
Shifflett Diana 508 Bainbridge Av 979-7035
Shifflett Doby & Patricia Rt6 286-4227
Shifflett Donk Ola Rt 621 974-7463
Shifflett James 2219 Williamsburg Rd
Shifflett James B 801 Stonehenge Av
Shifflett James C Stanardsville .....
Shifflett James E Earlysville .....
Shifflett James E Jr 552 Cleveland Av
Shifflett James F & Lois Longmeadow
Shifflett James F & Vernell Rt671 ...
Shifflett James J 1430 Rugby Av ....
Shifflett James K St George Av .....
Shifflett James L SR33 Stanardsville .
Shifflett James O Earlysville .....
Shifflett James O Stanardsville .....
Shifflett James R Old Lynchburg Rd ..
Shifflett James R Rt753 Esmont ....

Non-Comparison Sorts Don't Beat the Bound

Radix sort works by partitioning on the lowest order characters first, maintaining this order to break ties.

It takes time $O(nm)$ to sort n strings of length m, or time *linear* in the input size!

But m must be $\Omega(\log n)$ before the strings are all distinct!

Sorting n arbitrary, distinct keys cannot be done better than $\Theta(n \log n)$.