TIRGUL 6 in Data Structure – solution of last question – draft

(Remember that these notes are unchecked and spelling mistakes and inaccuracies may be plentiful)

Elad Verbin

April 28, 2006

1 Solution of last question from the TIRGUL

Question: You are given an array of size n with $\sqrt{n}$ distinct integers, each one appearing exactly $\sqrt{n}$ times. Prove a lower bound of $\Omega(n \log n)$ in the comparison model for any algorithm which sorts such an array.
Answer: First, assume that the $\sqrt{n}$ distinct values in the input array are $1, \ldots, \sqrt{n}$. This does not matter for an algorithm in the comparison model, because such an algorithm only performs comparisons between elements. Now, how many possible inputs are there? If all elements were distinct then the number of inputs would have been $n!$. However, in our case we need to divide by the number of permutations that will keep the array the same. Thus, the number of possible inputs is exactly

$$X = \frac{n!}{(\sqrt{n}!)^{\sqrt{n}}}$$

Now, let us look at some algorithm in the comparison model, that is a decision tree that performs only comparisons. How many leaves must it have? It is easily seen that a solution can be a correct solution only for one input array. Therefore the algorithm must have at least X leaves. It is a tree where every vertex has 2 or 3 children, and which has at least X leaves, so its depth is at least $\Omega(\log(X))$. Therefore, the running time of any algorithm in the comparison model is at least $\Omega(\log(X))$. Now we wish to prove that $\log(X) = \Omega(n \log n)$.

First, let us assume that for every m, $m! = 2^{m \log m}$ and in this case we will easily get that $\log(X) = \Omega(n \log n)$. The assumption that $m! = 2^{m \log m}$ is untrue, of course, but the math here is conceptually the right math, and it is cleaner than the correct proof we will give later. So:

$$\log(X) = \log\left(\frac{n!}{(\sqrt{n}!)^{\sqrt{n}}}\right) = \log(n!) - \log((\sqrt{n}!)^{\sqrt{n}}) = n \log n - \sqrt{n} \log(\sqrt{n}!) =$$
$$n \log n - \sqrt{n}\sqrt{n} \log(\sqrt{n}) = n \log n - n\frac{\log n}{2} = n \log n/2 = \Omega(n \log n) \ ,$$

as required.

Now, let us see that using the (correct) fact that $m! = 2^{\Theta(m \log m)}$ or, in other words, $\log(m!) =$

$\Theta(m \log m)$, does not give the required result:

$$\begin{aligned}\log(X) = \log\left(\frac{n!}{(\sqrt{n}!)^{\sqrt{n}}}\right) = \\ &\log(n!) - \log((\sqrt{n}!)^{\sqrt{n}}) = \Theta(n \log n) - \sqrt{n}\log(\sqrt{n}!) = \\ &\Theta(n \log n) - \sqrt{n}\Theta(\sqrt{n}\log(\sqrt{n})) = \Theta(n \log n) - \Theta(n \log n) = ??? \ .\end{aligned}$$

As you see, we can't get an answer here.[1]

So, how do we formally prove that $\log(X) = \Omega(n \log n)$ in a correct way? We use Stirling's approximation. Stirling's approximation states that

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \ .$$

In other words this says that

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} = 1 \ .$$

But we will be satisfied using the weaker, but still useful expression

$$n! = \Theta\left(\sqrt{n}\left(\frac{n}{e}\right)^n\right) \ .$$

From taking the logarithm of both sides of this expression it follows that

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

We now give a proof that $\log(X) = \Omega(n \log n)$ in a somewhat simpler way than how I proposed to do it in class:

$$\begin{aligned}\log(X) = \log\left(\frac{n!}{(\sqrt{n}!)^{\sqrt{n}}}\right) = \\ &\log(n!) - \log((\sqrt{n}!)^{\sqrt{n}}) = n \log n \pm O(n) - \sqrt{n}\log(\sqrt{n}!) = \\ &n \log n \pm O(n) - \sqrt{n}(\sqrt{n}\log(\sqrt{n}) \pm O(\sqrt{n})) = \\ &n \log n \pm O(n) - n\frac{\log n}{2} \pm O(n) = \Theta(n \log n) \pm O(n) = \Theta(n \log n) \ .\end{aligned}$$

2 Further Reading

I collected some web-pages which you might find interesting. If you take a look, I'm sure you'll find some interesting things.

2.1 The "P=NP?" Question

`http://en.wikipedia.org/wiki/Complexity_classes_P_and_NP`
`http://www.claymath.org/millennium/P_vs_NP/Official_Problem_Description.pdf`
`http://www.win.tue.nl/~gwoegi/P-versus-NP.htm`
`http://www.math.ias.edu/~avi/PUBLICATIONS/MYPAPERS/W06/w06.pdf`

[1] Specifically, computing a difference between Thetas is undefined, since it can have many answer: $0 = 10n - 10n = \Theta(n) - \Theta(n)$, but also $n = 11n - 10n = \Theta(n) - \Theta(n)$. This is much like the situation is Discrete Math, where the difference of two cardinalities (OTZMOT) is not defined: $\aleph - \aleph$ may be 0 or $\aleph$ or $\aleph_0$ or a finite number, depending on the choice of representatives.

2.2 Stirling's Approximation

http://hyperphysics.phy-astr.gsu.edu/Hbase/Math/stirling.html
http://en.wikipedia.org/wiki/Stirling's_approximation

2.3 About the 3-SUM problem

http://en.wikipedia.org/wiki/3SUM
http://geomblog.blogspot.com/2004/03/3sum.html

2.4 About lower bounds in computer science

Lecture notes L,H and J in http://compgeom.cs.uiuc.edu/~jeffe/teaching/373/