

Recursive formula for the average internal path length of a binary tree

The internal path length of a binary tree constructed from a random permutation of keys, as well as the recursive analysis of randomized quicksort, leads to the following expression for $P(n)$.

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

Instead of $\Theta(n)$ as used by Cormen/Leiserson/Rivest, we will use the more accurate expression $cn + d$ for constants c and d . It is not necessary to use d since this will be absorbed into cn by choosing c slightly larger, but our analysis of $P(n)$ will be completely precise in that way, so that the dependence on c and d becomes apparent.

Our goal is to get an exact formula for $P(n)$ based on

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

for $n > 1$, and known values for $P(0)$ and $P(1)$. In order to get rid of n in the denominator, multiply this equation by n :

$$nP(n) = 2 \sum_{i=0}^{n-1} P(i) + cn^2 + dn.$$

The next thing to get rid of is the sum. For that purpose, we write down the formula for $n + 1$ instead of n :

$$(n+1)P(n+1) = 2 \sum_{i=0}^n P(i) + c(n+1)^2 + d(n+1)$$

and subtract, which gets rid of the sum except for the last term in the sum, and of the terms cn^2 and dn :

$$(n+1)P(n+1) - nP(n) = 2P(n) + 2cn + c + d,$$

which is equivalent to

$$(n+1)P(n+1) = (n+2)P(n) + 2cn + c + d.$$

The left hand side has $(n+1)P(n+1)$ and the right hand side $(n+2)P(n)$, which can be turned into something that depends in the same way on n by dividing by $(n+1)(n+2)$. This yields the equivalent expression

$$\frac{P(n+1)}{n+2} = \frac{P(n)}{n+1} + \frac{2cn + c + d}{(n+1)(n+2)}$$

which, with

$$f(n) = \frac{P(n)}{n+1}, \quad g(n) = \frac{2cn + c + d}{(n+1)(n+2)},$$

we can write as

$$f(n+1) = f(n) + g(n).$$

Now we are nearly done, since this equation says that the term $f(n+1)$ is simply the previous term $f(n)$ plus a known expression $g(n)$. We simply use the equation repeatedly to see

$$\begin{aligned} f(n+1) &= f(n) + g(n) = f(n-1) + g(n-1) + g(n) \\ &= f(n-2) + g(n-2) + g(n-1) + g(n) \\ &= \dots = f(1) + g(1) + g(2) + \dots + g(n-1) + g(n). \end{aligned}$$

We will show that essentially, the known terms $g(i)$ are like $1/i$ times a constant, so that $g(1) + g(2) + \dots + g(n-1) + g(n) = \Theta(H_n)$ with the harmonic number H_n , defined by

$$H(n) = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n}.$$

More precisely, we have

$$g(n) = \frac{2c}{n+2} + \frac{d-c}{(n+1)(n+2)},$$

so that

$$g(1) + g(2) + \dots + g(n) = 2c(H_{n+2} - H_2) + (d-c)\left(\frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \dots + \frac{1}{(n+1)(n+2)}\right),$$

where

$$\frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \dots + \frac{1}{(n+1)(n+2)} = \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \frac{1}{4} + \dots + \frac{1}{n+1} - \frac{1}{n+2} = \frac{1}{2} - \frac{1}{n+2},$$

(this is called “telescoping” a sum) so that

$$g(1) + g(2) + \dots + g(n) = 2c(H_{n+2} - H_2) + (d-c)\left(\frac{1}{2} - \frac{1}{n+2}\right)$$

which is $\Theta(H_n)$ as claimed, since only constants or bounded terms are added to H_n . We could have used the Θ -notation earlier, in the form $g(n) = \Theta(1/n)$, for simplifying the derivation; you may feel more comfortable with the present reasoning.

Below, by comparing H_n with the integral of dt/t from 1 to n , it is shown that $H_n = \Theta(\log n)$. In summary, we get

$$\begin{aligned} P(n) &= (n+1) \cdot f(n) \\ &= (n+1) \cdot (f(1) + g(1) + g(2) + \dots + g(n-1)) \\ &= (n+1) \cdot (P(1)/2 + 2c(H_{n+1} - H_2) + (d-c)\left(\frac{1}{2} - \frac{1}{n+1}\right)) \\ &= \Theta(nH_n) = \Theta(n \log n). \end{aligned}$$

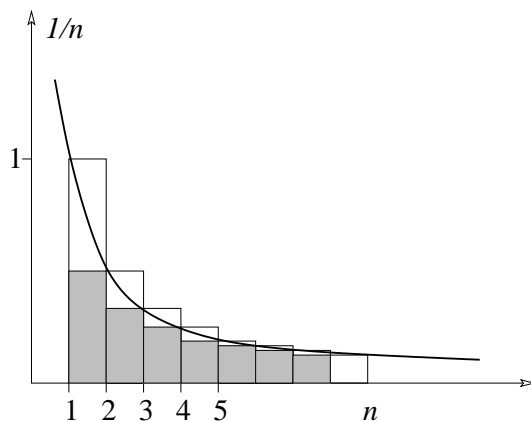
since the added constants and the term involving $\frac{1}{n+1}$ (which is at most $1/2$) do not matter.

Estimation of the harmonic number H_n

There is a very simple – but effective – method of estimating sums by integrals.¹ For estimating the *harmonic numbers*

$$H(n) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n},$$

we draw the following figure:



and derive from it

$$H_n - 1 = \sum_{k=2}^n \frac{1}{k} < \int_1^n \frac{1}{t} dt = \log n$$

by comparing the area below the graph of the function that maps t to $1/t$ for $1 \leq t \leq n$ with the area of the shaded rectangles, and

$$H_n - \frac{1}{n} = \sum_{k=1}^{n-1} \frac{1}{k} > \int_1^n \frac{1}{t} dt = \log n$$

by comparing the integral with the area of the large rectangles (including the shaded parts). Taken together, this yields

$$\log n + \frac{1}{n} < H_n < \log n + 1.$$

In particular, $\lim_{n \rightarrow \infty} H_n = \infty$, and the order of growth of H_n is given by $\lim_{n \rightarrow \infty} \frac{H_n}{\log n} = 1$.

This is an even better estimate than $H_n = \Theta(\log n)$, since the constant factor for comparing the two functions (of n) H_n and $\log n$ is arbitrarily close to 1.

¹The following paragraphs are adapted from *Proofs from the Book*, M. Aigner and G. M. Ziegler, Springer-Verlag, 2003.