An analysis of randomized Quicksort
CSCE 750, Fall 2012

Stephen A. Fenner

October 2, 2012

We showed in class that the expected run time $E(n)$ of randomized Quicksort run on a list of n items asymptotically satisfies the recurrence

$$E(n) = n + \frac{2}{n}\sum_{k=0}^{n-1} E(k)\ .$$

Here we show via the substitution method that $E(n) = O(n \lg n)$. We prove by induction on n that $E(n) \le cn \lg n$ for some sufficiently large constant c chosen later. We give the inductive step, then afterward describe how the base cases can be handled. The key trick of the proof is to split the sum into *two* pieces.

Fix n sufficiently large. We assume (inductive hypothesis) that $E(m) \le cm \lg m$ for all $0 \le m < n$. Here we "imagine" that $0 \lg 0 = 0$, so we start the sum with $k = 1$ after we use the inductive hypothesis. This is not really legitimate, of course, so we show how to fix it afterwards. We have

$$\begin{aligned}
E(n) &= n + \frac{2}{n}\sum_{k=0}^{n-1} E(k) \le n + \frac{2}{n}\sum_{k=1}^{n-1} ck \lg k && \text{(inductive hypothesis)}\\
&= n + \frac{2c}{n}\left(\sum_{k=1}^{\lfloor n/2\rfloor} k \lg k + \sum_{k=\lfloor n/2\rfloor+1}^{n-1} k \lg k\right)\\
&\le n + \frac{2c}{n}\left(\sum_{k=1}^{\lfloor n/2\rfloor} k \lg \frac{n}{2} + \sum_{k=\lfloor n/2\rfloor+1}^{n-1} k \lg n\right)\\
&= n + \frac{2c}{n}\left((\lg n - 1)\sum_{k=1}^{\lfloor n/2\rfloor} k + \lg n \sum_{k=\lfloor n/2\rfloor+1}^{n-1} k\right)\\
&= n + \frac{2c}{n}\left(\lg n \sum_{k=1}^{n-1} k - \sum_{k=1}^{\lfloor n/2\rfloor} k\right)\\
&\le n + \frac{2c}{n}\left(\frac{n^2 \lg n}{2} - \frac{\lfloor n/2\rfloor(\lfloor n/2\rfloor - 1)}{2}\right)\\
&\le n + \frac{2c}{n}\left(\frac{n^2 \lg n}{2} - \frac{(n/3)^2}{2}\right) = cn \lg n - \frac{cn}{9} + n \le cn \lg n
\end{aligned}$$

provided $c \ge 9$.

Handling the base cases

We really cannot assume that $E(m) \leq cm \lg m$ for all $0 \leq m < n$, because that would give $E(1) \leq c \lg 1 = 0$, and worse, $0 \lg 0$ is not even well-defined. However, we *can* start the induction at $m = 2$, that is, we can assume that $E(2) \leq 2c \lg 2 = 2c$ by choosing c large enough (and similarly for any fixed finite set of $E(m)$ with $m \geq 2$). We still have $E(0)$ and $E(1)$ terms appearing in the sum, though, and these are positive values. So let $d = \max(E(0), E(1), 1)$. We can then use the following recurrence for all $n \geq 2$:

$$E(n) = n + \frac{2}{n}\sum_{k=0}^{n-1} E(k) \leq n + \frac{2}{n}\left(2d + \sum_{k=2}^{n-1} E(k)\right) \leq n + 2d + \frac{2}{n}\sum_{k=2}^{n-1} E(k) \leq 2dn + \frac{2}{n}\sum_{k=2}^{n-1} E(k) \; .$$

That is,

$$E(n) \leq 2dn + \frac{2}{n}\sum_{k=2}^{n-1} E(k) \; .$$

The inductive step now works legitimately with this recurrence, because we only need to assume as our inductive hypothesis that $E(m) \leq cm \lg m$ for $2 \leq m < n$, which we can do. As before, we get $E(n) = O(n \lg n)$. This new recurrence just serves to boost the value of $E(n)$ by a factor of $2d$ (at most), so this does not change the asymptotic behavior of $E(n)$.