Average Case Analysis of Quicksort

Let $T_{avg}(n)$ be the average time taken by quicksort on an array of n distinct elements. In lecture, we saw recurrence

$T_{avg}(n) = \frac{1}{n}[\sum_{j=1}^{n} T_{avg}(j-1) + T_{avg}(n-j)] + dn$, for some $d > 0$.

Adding the base case, we get

$$T_{avg}(n) = \begin{cases} 1 & n \leq 1 \\ \frac{1}{n}[\sum_{j=1}^{n} T_{avg}(j-1) + T_{avg}(n-j)] + dn & n > 1 \end{cases}$$

In the above summation as j varies,

- $T_{avg}(j-1)$ ranges from $T_{avg}(0)$ to $T_{avg}(n-1)$ and
- $T_{avg}(n-j)$ ranges from $T_{avg}(n-1)$ to $T_{avg}(0)$

$$\Rightarrow T_{avg}(n) = \begin{cases} 1 & n \leq 1 \\ \frac{2}{n}[\sum_{l=0}^{n-1} T_{avg}(l)] + dn & n > 1 \end{cases}$$

$$\Rightarrow T_{avg}(n) \leq \begin{cases} d_0 & n \leq 3 \\ \frac{2}{n}[\sum_{l=0}^{n-1} T_{avg}(l)] + dn & n > 3 \end{cases} \cdots (I) \qquad \text{[By taking } d_0 = max\{1, T(2), T(3)\}\text{]},$$

Using (I), we can prove the following.

Claim: $T_{avg}(n) \leq \begin{cases} d_0 & n \leq 3 \\ c\,n \ln n & n > 3 \end{cases},$

where $c = 4(d + 4d_0)$.

[Note that this immediately implies that $T_{avg}(n)$ is $O(n \log n)$.]

Proof (of Claim): We prove this by induction on n.

Base Case: $n = 0$. Follows by definition.

Induction Step: Let the claim hold for all $n < k$. We show it for $n = k$. Case of $k \leq 3$ is trivial. So, let $k > 3$.

By (I),

$T_{avg}(k)$

$\leq \frac{2}{k}[\sum_{l=0}^{k-1} T_{avg}(l)] + dk$

$= \frac{2}{k}[\sum_{l=0}^{3} T_{avg}(l) + \sum_{l=4}^{k-1} T_{avg}(l)] + dk$

$\leq \frac{2}{k}[4d_0 + \sum_{l=4}^{k-1} T_{avg}(l)] + dk$

$\leq \frac{2}{k}[4d_0 + \sum_{l=4}^{k-1} c\,l \ln l] + dk$ (by induction hypothesis)

$\leq \frac{2}{k}[4d_0 + c \int_4^k l \ln l \, dl] + dk$

$\leq \frac{8d_0}{k} + \frac{2c}{k} \int_4^k l \ln l \, dl + dk$

$\leq 8d_0 + \frac{2c}{k} \int_4^k l \ln l \, dl + dk$

It can be shown (for example, using integration by parts)

$\int l \ln l \, dl = \frac{l^2}{4}[2 \ln l - 1]$

$\Rightarrow \int_4^k l \ln l \, dl$

$= \frac{k^2}{4}[2 \ln k - 1] - 4(2 \ln 4 - 1)$

$\leq \frac{k^2}{4}[2 \ln k - 1]$ (because $2 \ln 4 - 1 > 0$)

$$
\begin{aligned}
&\Rightarrow T_{avg}(n)\\
&\leq 8d_0 + \frac{2c}{k} \times \frac{k^2}{4}[2\ln k - 1] + dk\\
&= 8d_0 + \frac{ck}{2}[2\ln k - 1] + dk\\
&= c\,k\ln k + 8d_0 - \frac{ck}{2} + dk\\
&= c\,k\ln k + 8d_0 - 2(d + 4d_0)k + dk \qquad \text{(substituting value of } c \text{ in } \frac{ck}{2})\\
&= c\,k\ln k + 8d_0(1-k) - dk\\
&\leq c\,k\ln k + 8d_0(1-k) \text{ (as } d, k > 0)\\
&\leq c\,k\ln k \text{ (as } d_0 > 0 \text{ and } k > 1)
\end{aligned}
$$

This completes the induction step. □

We have proved that $T_{avg}(n)$ is $O(n \log n)$.

A question remains, how did we choose c and how did we decide to break the cases into $n \leq 3$ and $n > 3$. This is done by letting c, n_0 as variables and writing the induction step in terms of them. One may then choose the values for these variables so that constraints implies by induction step holds.