BLACKBOARD PROOFS

CSE202 – WEEK 6

1. Average-case complexity of Quicksort

The starting point is to decompose the expected number of comparisons according to the location of the pivot:

$$\begin{aligned}\mathbb{E}(C_n) &= \sum_{i=1}^{n} \mathbb{E}(C_n \mid \text{pivot at } i)\Pr(\text{pivot at } i)\\ &= \sum_{i=1}^{n} \mathbb{E}(n-1+C_{i-1}+C_{n-i})\frac{1}{n},\end{aligned}$$

whence by linearity of expectation

$$E_n = n-1+\sum_{i=1}^{n}\frac{E_{i-1}+E_{n-i}}{n}. \tag{1}$$

By reorganization of the final sum, this becomes

$$E_n = n-1+\frac{2}{n}(E_0+\cdots+E_{n-1}).$$

The sum will be disposed of by first isolating it

$$nE_n - n(n-1) = 2(E_0+\cdots+E_{n-1})$$

and subtracting two consecutive values, which leads to a simple linear recurrence

$$nE_n-(n-1)E_{n-1}-2(n-1)=2E_{n-1}.$$

Rearranging and dividing by $n(n+1)$ gives

$$\frac{E_n}{n+1}-\frac{E_{n-1}}{n}=\frac{2(n-1)}{n(n+1)}=\frac{4}{n+1}-\frac{2}{n}. \tag{2}$$

Now, the left-hand side telescopes when summing for $n=1$ to N, leading to

$$\frac{E_N}{N+1}=4\left(\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{N+1}\right)-2\left(1+\frac{1}{2}+\cdots+\frac{1}{N}\right)=2H_N-\frac{4N}{N+1},$$

where

$$H_N = 1+\cdots+\frac{1}{N}=\log N+\gamma+O(1/N),\quad N\to\infty,$$

is the Nth *harmonic number*, and $\gamma\approx 0.577$ denotes Euler's constant.

Note on the existence of $\gamma=\lim u_n$, where $u_n=H_n-\log n$. This can be seen by observing that the sequence u_n is decreasing and bounded from below. It is decreasing:

$$u_n-u_{n-1}=\frac{1}{n}+\log\frac{n-1}{n}\le\frac{1}{n}-\frac{1}{n}=0,$$

where the inequality comes from $\log(1+u) \leq u$ for all $u > -1$. It is also bounded below, since

$$\sum_{i=1}^{n} \frac{1}{i} > \sum_{i=1}^{n-1} \int_{i}^{i+1} \frac{dt}{t} = \int_{1}^{n} \frac{dt}{t} = \log n.$$

2. A Self-Contained Derivation of the Variance of the Number of Comparisons in Quicksort (Read only if you care.)

Let $Q_n(u)$ be the sum

$$Q_n(u) = \sum_{k \geq 0} p_{n,k} u^k,$$

with $p_{n,k}$ the probability that Quicksort performs k comparisons when sorting n elements. The derivative of $Q_n(u)$ evaluated at 1 recovers the average number of comparisons, as

$$Q_n'(1) = \sum_{k \geq 0} k p_{n,k} = \mathbb{E}(C_n) = E_n.$$

Similarly, the variance can be deduced from

$$Q_n''(1) = \sum_{k \geq 0} k(k-1) p_{n,k} = \mathbb{E}(C_n^2 - C_n) = \mathbb{E}(C_n^2) - \mathbb{E}(C_n),$$

so that the variance is given by

$$V_n = \mathbb{E}(C_n^2) - (\mathbb{E}(C_n))^2 = Q_n''(1) + Q_n'(1) - Q_n'(1)^2.$$

Our goal is to obtain sufficient control over $Q_n(u)$ so these quantities follow easily.

From the Quicksort recurrence

$$C_n = n - 1 + C_{i-1} + C_{n-i}$$

when the pivot is at index i, which happens with probability $1/n$, we deduce that

$$p_{n,k} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{k-(n-1)} p_{i-1,j} p_{n-i,k-(n-1)-j},$$

which expresses that fact that if the total number of comparisons is k and the first recursive call uses j comparisons, then the second one uses $k-(n-1)-j$ comparisons, since the first partitioning uses $n-1$ comparisons.

Multiplying by u^k and summing over all possible values of k gives

$$\begin{aligned} Q_n(u) &= \sum_{k \geq 0} \frac{u^k}{n} \sum_{i=1}^{n} \sum_{j=0}^{k-(n-1)} p_{i-1,j} p_{n-i,k-(n-1)-j} \\ &= \frac{u^{n-1}}{n} \sum_{i=1}^{n} \sum_{k \geq 0} \sum_{j=0}^{k-(n-1)} (p_{i-1,j} u^j)(p_{n-i,k-(n-1)-j} u^{k-(n-1)-j}) \end{aligned}$$

whence, by recognizing a product of polynomials,

$$Q_n(u) = \frac{u^{n-1}}{n} \sum_{i=1}^{n} Q_{i-1}(u) Q_{n-i}(u). \tag{3}$$

This can be used to compute the first few polynomials:

$$Q_1 = 1, \quad Q_2 = u, \quad Q_3 = \frac{1}{3}u^2 + \frac{2}{3}u^3, \quad Q_4 = \frac{1}{2}u^4 + \frac{1}{6}u^5 + \frac{1}{3}u^6.$$

For instance, the last one means that when sorting 4 elements, the algorithm uses 5 comparisons with probability 1/6.

For all k, $Q_k(1) = 1$ since it is the sum of all possible probabilities. Thus, differentiating and evaluating the previous equation at 1 gives

$$Q'_n(1) = n - 1 + \frac{1}{n}\sum_{i=1}^{n}(Q'_{i-1}(1) + Q'_{n-i}(1)),$$

which is exactly Eq. (1).

For this polynomial to really become useful, we go one step further and introduce the power series with polynomial coefficients

$$Q(z,u) = \sum_{n\geq 0} Q_n(u)z^n = 1 + z + uz^2 + \left(\frac{1}{3}u^2 + \frac{2}{3}u^3\right)z^3 + \cdots.$$

Multiplying both sides of Eq. (3) by nuz^n and summing over n gives

$$\sum_{n\geq 0} nuQ_n(u)z^n = uz\sum_{n\geq 0}\sum_{i=1}^{n}(Q_{i-1}(u)(uz)^{i-1})(Q_{n-i}(u)(uz)^{n-i})$$
$$uz\frac{\partial Q(z,u)}{\partial z} = uzQ(zu,u)^2,$$

which simplifies to

$$\frac{\partial Q(z,u)}{\partial z} = Q(zu,u)^2, \quad \text{with } Q(0,u) = 1. \tag{4}$$

This might not look so hopeful, but by evaluating at $u = 1$ this equation and its derivative with respect to u, we get

$$\frac{\partial Q}{\partial z}(z,1) = Q^2(z,1), \quad \frac{\partial^2 Q}{\partial z \partial u}(z,1) = 2Q(z,1)\left(z\frac{\partial Q}{\partial z}(z,1) + \frac{\partial Q}{\partial u}(z,1)\right).$$

The first equation has for solution $Q(z) = 1/(1-z)$, which is nothing but the fact that all $Q_n(1)$ equal 1. Thus the second one is a linear equation in $q_1(z) = \partial Q/\partial u(z,1)$:

$$q'_1(z) = \frac{2}{1-z}\left(\frac{z}{(1-z)^2} + q_1(z)\right), \qquad q'_1(0) = 0$$

of which the solution can be checked to be

$$q_1(z) = \frac{1}{(1-z)^2}\left(2\ln\frac{1}{1-z} - 2z\right),$$

whose Taylor expansion has for coefficients precisely the expectations

$$E_n = 2(n+1)H_n - 4n$$

from the previous section.

Differentiating Eq. (4) once more with respect to u and evaluating at $u = 1$ gives a linear differential equation for $q_2(z) = \partial^2 Q/\partial u^2(z,1)$:

$$(1-z)^4 q'_2(z) - 2(1-z)^3 q_2(z) = 8\ln^2\frac{1}{1-z} - 8z\ln\frac{1}{1-z} - 2z^2,$$

of which the solution with $q_2(0) = 0$ can be seen to be

$$q_2(z) = 4\frac{1+z}{(1-z)^3}\ln^2\frac{1}{1-z} - 4\frac{1+z}{(1-z)^3}\ln\frac{1}{1-z} + \frac{2z(2+z)}{(1-z)^3}.$$

From there, an explicit formula for $Q_n''(1)$ follows by extracting the coefficient of z^n in the Taylor expansion, leading to

$$Q_n''(1) = 4(n+1)^2 H_n^2 - 4(n+1)(4n+1)H_n - 4(n+1)^2 H_n^{(2)} + n(23n+17),$$

where

$$H_n^{(2)} = \sum_{i=1}^{n} \frac{1}{i^2} = \frac{\pi^2}{6} + O(1/n), \quad n \to \infty.$$

The general formula for the variance is thus obtained as

$$V_n = Q_n''(1) + Q_n'(1) - Q_n'(1)^2 = n(7n+13) - 4(n+1)^2 H_n^{(2)} - 2(n+1)H_n.$$

The asymptotic behaviour follows:

$$V_n \sim (7 - 2\pi^2/3)n^2, \quad n \to \infty$$

and thus the standard deviation $\approx 1.93n$ becomes negligible compared to the mean $\sim 2n \log n$ as $n \to \infty$.