EXERCISE FOR CSE202 – WEEK 11

The aim of this exercise is to estimate the average-case complexity of the last-character heuristic used in the Boyer-Moore algorithm. The hypotheses are that the pattern has length m smaller than the size R of the alphabet and that the text is a sequence of n letters drawn independently and uniformly at random from the alphabet.

Question 1. *Show that the expected length of a shift is at least*

$$\ell_m := m\left(1-\frac{m}{R}\right)+\sum_{i=1}^{m-1}\frac{i}{R} = m\left(1-\frac{m+1}{2R}\right).$$

Solution. Let ℓ be the letter of the text compared with the last letter of the pattern. Its value is any letter of the alphabet with probability $1/R$. Let $\ell_1,\dots,\ell_k$ be the distinct letters in the pattern ($k \le m$). For each of them, the shift s_{ℓ_i} performed when ℓ is compared to the letter ℓ_i is to the rightmost occurrence of ℓ_i in the pattern (as we are computing a lower bound, we can consider a shift 0 when ℓ is the last letter of the pattern). So the expectation is at least

$$\mathbb{P}(\ell \notin \{\ell_1,\dots,\ell_k\})m + \sum_{i=1}^{k}\mathbb{P}(\ell=\ell_i)s_{\ell_i}.$$

As the shifts s_{ℓ_i} are distinct integers in $\{0,\dots,m-1\}$, the second summand is lower bounded by $\sum_{i=0}^{m-1} i/R$. Finally, since $k \le m$, the first probability is lower bounded by $1-m/R$. Putting these bounds together yields the desired lower bound. □

Question 2. *Consider an infinite text, a random positive integer N and let $S_1,\dots,S_N$ be the lengths of the first N (non-independent) shifts performed during a search for the pattern in the text. Show that*

$$\mathbb{E}\left(\sum_{i=1}^{N} S_i\right) \ge \ell_m \mathbb{E}(N).$$

[Indication: the variables S_i are independent from N.]

Solution. Using the indication, the expectation of the sum is

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{N} S_i\right) &= \sum_{k\ge 1}\mathbb{P}(N=k)\mathbb{E}\left(\sum_{i=1}^{k} S_i \,\middle|\, N=k\right)\\ &= \sum_{k\ge1}\mathbb{P}(N=k)\sum_{i=1}^{k}\mathbb{E}(S_i|N=k) = \sum_{k\ge1}\mathbb{P}(N=k)\sum_{i=1}^{k}\mathbb{E}(S_i)\\ &\ge \ell_m\sum_{k\ge1}k\mathbb{P}(N=k) = \ell_m\mathbb{E}(N). \quad \square\end{aligned}$$

Question 3. *Show that the expected number of shifts when reading a text of length n is $\le n/\ell_m$.*

Solution. Stopping the search when $n-m$ characters have been visited leads to the inequality

$$n \geq \mathbb{E}\left(\sum_{i=1}^{N} S_i\right) \geq \ell_m \mathbb{E}(N),$$

where the last inequality comes from the previous question. Thus $\mathbb{E}(N) \leq n/\ell_m$, as was to be proved. □

Question 4. *Show that the expected number of comparisons made before performing a shift is at most*

$$c_m := 1 + \frac{m-1}{R}.$$

Solution. One comparison is always needed for the last character. With probability $1-1/R$ it results in a shift. Otherwise, at most $m-1$ more comparisons will be performed before the next shift. The result follows from

$$\left(1-\frac{1}{R}\right) \times 1 + \frac{1}{R} \times m = 1 + \frac{m-1}{R}.$$ □

Question 5. *Show that the expected number of comparisons of the last character heuristic for the whole text is at most*

$$\frac{n}{m}\frac{1+\frac{m-1}{R}}{1-\frac{m+1}{2R}} = \frac{n}{m}\left(\frac{6R-4}{2R-m-1} - 2\right).$$

[Indication: proceed as in Question 2.]

Solution. Let $C_1, \ldots, C_N$ be the number of comparisons made before each shift. These variables are independent of N. Then the total number of comparisons has an expectation

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{N} C_i\right) &= \sum_{k\geq 1} \mathbb{P}(N=k) \sum_{i=1}^{k} \mathbb{E}(C_i | N=k) = \sum_{k\geq 1} \mathbb{P}(N=k) \sum_{i=1}^{k} \mathbb{E}(C_i) \\ &\leq c_m \sum_{k\geq 1} k\mathbb{P}(N=k) = c_m \mathbb{E}(N).\end{aligned}$$

By the previous results this is bounded by nc_m/ℓ_m and the result is obtained by injecting their values. □

Question 6. *Show that for $m < R$, the expected number of comparisons is smaller than $4n/m$.*

Solution. The function $(6R-4)/(2R-m-1)$ is increasing with m for $m < 2R-1$, hence its maximal value for $m < R$ is reached at $m = R-1$ where its value is $6 - 4/R < 6$. □

Question 7. *Finally, show that as $R/m \to \infty$ (ie, when the alphabet is very large), the expected number of comparisons is equivalent to n/m.*

Solution. The result of the Question 5 behaves like

$$\frac{n}{m}\left(1 + O\left(\frac{m}{R}\right)\right) \sim \frac{n}{m}.$$

This is an upper bound on the expected number of comparisons. The conclusion comes from the fact that this value is also a lower bound on the number of comparisons. Indeed, if fewer than n/m characters are checked, then there is a sequence of length m that has not been checked and could contain the pattern. □