Lecture - 16,17

Thursday, 1 September 2016 (14:25 - 15:15; 17:10 - 18:30)

Longest Streak problem

Assume that we toss a fair coin n times and generate a binary string accordingly : if the toss results in a Head we note a 0 and a 1 if we get a Tail. This way we obtain a binary string of length n.

Definition 1. *A streak is defined as the sub-string of all consecutive 0's. The length of a streak is given by the number of 0's that make up the streak.*

Question : Given a string of length n generated by the above experiment, what is the expected length of the longest streak?

The aim of this lecture was to show that, if s is the expected length of the longest streak, then we would show :

$$s = \Omega(logn) \text{ and } s = O(logn)$$
$$\therefore s = \Theta(logn)$$

Observation 1 :

Firstly, we will provide an upper bound on the expected length of the longest streak.
We begin by defining a random variable L to denote the length of the longest streak in a string of length n. Let L_j be the event that we get a string where the longest streak length is precisely j. Therefore, by the definition of expectation, it is $E[L]$ that we desire to compute which can be written as :

$$E[L] = \sum_{j=1}^{n} j.Pr(L_j)$$

We define another event $A_{i,k}$ as the obtaining a string which has a streak of length k starting from the i^{th} index of the given string.
Thus, we can say that $Pr(A_{i,k}) = \frac{1}{2^k}$. When we assume the streak length to be $2logn$ we see that :

$$Pr(A_{i,k}) = Pr(A_{i,2logn}) = \frac{1}{2^{2logn}} = \frac{1}{n^2} \tag{1}$$

Now, we can compute the probability that we obtain a streak of length $k = 2logn$ at some point in the string as :

$$\begin{aligned} Pr\left(\bigcup_{i=1}^{n-k} A_{i,k}\right) &\leq \sum_{i=1}^{n-k} Pr(A_{i,k}) && \because \text{Union Bound} \\ &= \sum_{i=1}^{n-k} \frac{1}{n^2} && \because \text{of Eqn. (1)} \\ &\leq \frac{1}{n} \end{aligned}$$

We shall bear in mind the above result and use it later on. Let is reconsider the expected value of longest streak length.

$$
\begin{aligned}
E[L] &= \sum_{j=1}^{n} j.Pr(L_j) \\
&= \sum_{j=1}^{2logn-1} j.Pr(L_j) + \sum_{j=2logn}^{n} j.Pr(L_j) \\
&\leq \left(2log(n) \sum_{j=1}^{2\ log(n)} Pr(L_j) \right) + \left(n \sum_{j=2logn}^{n} Pr(L_j) \right) \\
&\leq 2log(n).1 + n.\frac{1}{n}
\end{aligned}
$$

The last inequality holds from the fact that:

$$\sum_{j=1}^{n} Pr(L_j) = 1$$

$$\therefore \sum_{j=1}^{2logn} Pr(L_j) \leq 1$$

Additionally, the probability of finding a string having the longest streak of length $2logn$ or more is denoted by :

$$\sum_{j=2logn}^{n} Pr(L_j) \ = Pr\left(\bigcup_{i=1}^{n-k} A_{i,k} \right) \leq \frac{1}{n}$$

from the previous definition.

Thus, we have shown that the expected length of the longest streak is $E[L] = O(log\ n)$, i.e the expected length is bounded above by order $log\ n$.

Now, we will work towards showing that in fact $E[L]$ is bounded below by $\Omega(log\ n)$.

Let us assume the string to be broken into slabs of length s, such that we have the string partitioned into a total of $\frac{n}{s}$ slabs.

The probability that a streak is present in a given slab $= \frac{1}{2^s}$.

Let us assume the slab size to be $s = \frac{log\ n}{2}$. Therefore, the probability that a streak is present in a given slab $= \frac{1}{2^{\frac{log\ n}{2}}} = \frac{1}{\sqrt[2]{n}}$.

The probability that a streak is not present in a given slab $= 1 - \frac{1}{\sqrt{n}}$.

The probability that a streak is not present in any of the slabs of the given string $= \left(1 - \frac{1}{\sqrt{n}}\right)^{\frac{2n}{log\ n}}$

$$= \left(1 - \frac{1}{\sqrt{n}}\right)^{\frac{2.\sqrt{n}.\sqrt{n}}{log\ n}}$$

$$\approx e^{\frac{-2\sqrt{n}}{log\ n}}$$

$$\geq \frac{1}{e^{log\ n}}$$

$$= \frac{1}{n}$$

The last inequality follows from the fact that $\frac{2\sqrt{n}}{log\ n} \geq log\ n$. Therefore, we can say that a streak of length $\frac{log\ n}{2}$ is present in at least one of the slabs = 1 - O($\frac{1}{n}$). Since we are considering streaks only within a slab, we are discounting on streaks that could lie across slabs. Hence, the probability that there is a streak of length at least $\frac{log\ n}{2}$ anywhere in the string is greater than the above probability. Thus

$$\sum_{j=\frac{logn}{2}}^{n} Pr(L_j) \geq 1 - O\left(\frac{1}{n}\right)$$

Thus, we have:

$$E[L] = \sum_{j=1}^{\frac{log\ n-1}{2}} j.Pr(L_j) + \sum_{j=\frac{logn}{2}}^{n} j.Pr(L_j)$$

$$\geq \sum_{j=\frac{logn}{2}}^{n} j.Pr(L_j)$$

$$\geq \frac{log\ n}{2}\left(1 - O\left(\frac{1}{n}\right)\right)$$

$$\geq \frac{log\ n}{2}$$

This allows us to conclude that the expected length of longest streak is also bounded below by $O(log\ n)$.