Probab. Th. Rel. Fields 87, 313–323 (1991)

Probability Theory and Related Fields

On the waiting time till each of some given patterns occurs as a run

Tamás F. Móri
Department of Probability Theory, Eötvös University, Múzeum krt. 6-8,
H-1088 Budapest, Hungary

Received September 19, 1986; in revised form April 15, 1990

Summary. A limit theorem is proved for the waiting time till each of a given set of length n patterns occurs as a run in a sequence of i.i.d. random variables distributed uniformly on $\{1, 2, \ldots, d\}$. A heuristic approach called the independence principle is introduced which can be applied to similar problems connected with waiting times.

1. Introduction

Let H be a finite alphabet; we can suppose $H = \{1, 2, \ldots, d\}$. Let H^n denote the set of words of length n over H. Consider a sequence $X_1, X_2, \ldots$ of i.i.d. random variables distributed uniformly on H. For any word $A \in H^n$ define

$$T(A) = \inf\{m: (X_{m-n+1} X_{m-n+2} \ldots X_m) \equiv A\} ,$$

the waiting time till A appears as a run in the sequence of repeated experiments.

Problems connected with runs and waiting times are very popular in the classical theory of probability, for they can be formulated without difficult notions or involved technical terms, their solutions, however, are far from being trivial, and they help to understand the nature of randomness (Erdös and Rényi (1970) or Erdös and Révész (1976)). They are also appealing for combinatorists (Guibas and Odlyzko 1981). Recently this topic is being applied at an increasing rate in several fields such as computer science (complexity theory), Monte Carlo methods and simulation (generating pseudorandom numbers, tests for randomness) or molecular biology (statistical analysis of long DNA and RNA molecules). These applications constitute a constant source of inspiration and orientation for the research (Shukle and Shrivastava 1985).

The aim of the present paper is to find the limit distribution of the maximum of several waiting times. Given a subset $H_n \subset H^n$ we are interested in the number of experiments needed till each word from H_n occurs as a run at least once. Let $W_n = \max\{T(A): A \in H_n\}$. Our main result is the following theorem, to be proved in Sect. 3.

Theorem 1. *Suppose*

$$\lim_{n\to\infty} |H_n| = +\infty \ .$$

Then

$$\lim_{n\to\infty} P(d^{-n}W_n - \log|H_n| \leqq y) = e^{-e^{-y}}$$

for every real y.

2. Independence principle: a heuristic approach

We now attempt to enlighten the background of Theorem 1 by presenting an intuitive method which can often be applied to problems connected with the waiting times $T(A)$.

A most simplified description of the joint distribution of these waiting times, which we shall call *the independence principle* says that they can be regarded as independent, exponentially distributed random variables with common expectation d^n, unless something tells against it. It is easy to see that Theorem 1 lies within the scope of this principle. Let us see its motivation as well as the limits of its applicability.

Concerning a single $T(A)$ it is not so difficult to see the approximate exponentiality of its distribution. This was done in Móri and Székely (1984), together with some extensions to joint distributions in a special case. Their method of proof was based on the characteristic "lack of memory" property of the exponential distribution. This method was refined in Móri (1985) in order to obtain general estimations for the joint distribution of several $T(A)$'s. It turned out that the joint asymptotic distribution of a couple of waiting times had independent exponential marginals as far as the pairwise overlapping between the words was negligible, otherwise the dependence was of Marshall-Olkin type. The results were formulated to cover large deviations, too. These papers can be considered as the mathematical justification of the heuristic independence principle.

Let us recall some details which we shall need in the sequel.

Lemma 1 (Móri 1985). *For* $A_1, A_2, \ldots, A_r \in H^n$ *let* $Z = \min T(A_i)$ *and* $b = nrd^{-n} < \frac{1}{5}$. *Then for every* $y > 0$

$$\exp\left(-\frac{1-4b}{1-5b}y\right) \leqq P\left(\frac{Z}{E(Z)} > y\right) \leqq \frac{1-4b}{1-5b}e^{-y} \ .$$

In order to apply this lemma we must know how to calculate $E(Z)$, the expected number of experiments needed till any of r competing words appears as a run. This expectation is provided by the so called "magic" Conway algorithm. In his paper (1980) Li gave an elegant proof to a generalization of the Conway algorithm. This algorithm defines a measure of overlapping between two words $A = (a_1 a_2 \ldots a_n)$ and $B = (b_1 b_2 \ldots b_n)$ as $A*B = \sum_{i=1}^{n} \varepsilon_i d^i$, where $\varepsilon_i = 1$ if the words $(a_{n-i+1} \ldots a_n)$ and $(b_1 \ldots b_i)$ are identical, otherwise $\varepsilon_i = 0$. $A*B$ is called the *leading number* of A over B.

Lemma 2 [Li (1980)]. *Let $A_1, \ldots, A_r$ and Z as above and let $p_i = P(T(A_i) = Z)$, the probability that A_i is the first to appear. Then for every $j = 1, 2, \ldots, r$*

$$\sum_{i=1}^{r} p_i A_i * A_j = E(Z) .$$

That is, $p_1, \ldots, p_r$ and $E(Z)$ are the solution of a system of $r+1$ linear equations (the $(r+1)$st is that $p_1 + \cdots + p_r = 1$). In particular, when $r = 1$, we have $E(T(A)) = A * A$.

When the overlapping is negligible, $A_i * A_j \approx 0$ $(i \neq j)$, thus the j'th equation gives $p_j E(T(A_j)) \approx E(Z)$. Hence

$$E(Z)^{-1} \approx \sum_{i=1}^{r} E(T(A_i))^{-1} , \tag{1}$$

which is in accordance with the approximate independence of the waiting times $T(A_i)$. In fact, the regular behaviour of the minima is equivalent to that of the waiting times themselves.

Let us see a counterexample, where the independence principle does not work.

If, instead of the maximum, one asks the minimum of $T(A)$ as A runs over H_n, for $H_n = H^n$ the independence principle says that it is of exponential distribution with expectation 1, but obviously it is identical with n. The irregular behaviour of the minimum waiting time cannot be excluded by imposing restrictions on the growth of $|H_n|$. Of course, there are regular subsets (in the sense of (1)); it can be shown by the help of Lemma 2 that most subsets of H^n are regular. However, for any r, $2 \leqq r \leqq d^n$, one can find a set $H_n \subset H^n$, $|H_n| = r$, such that

$$E(Z)^{-1} \leqq \frac{2d+3}{2d+4} \sum_{A \in H_n} E(T(A))^{-1} . \tag{2}$$

For the construction consider first a directed graph with set of vertices H^{n-1} and set of edges H^n. Edge $(a_1 \ldots a_n)$ points from vertex $(a_1 \ldots a_{n-1})$ at vertex $(a_2 \ldots a_n)$. Then every vertex has indegree d and outdegree d, thus the graph is Eulerian, that is, the length n words over H can be arranged in a cycle C in such a way that every word is overlapped by its cyclic successor in $n-1$ letters. For any given r, $2 \leqq r \leqq d^n$, let H_n consist of r consecutive in C words: $H_n = \{A_1, A_2, \ldots, A_r\}$. We can suppose that $E(T(A_r)) \geqq E(T(A_i))$, $1 \leqq i < r$. If A_i overlaps itself in length $l > 1$, then it also overlaps A_{i-1} in length $l-1$. Hence $A_i * A_i \leqq d + dA_{i-1} * A_i$ if $A_i * A_i \geqq d^n + d$, and $A_i * A_i \leqq dA_{i-1} * A_i$ if $A_i * A_i = d^n$. In both cases

$$A_{i-1} * A_i \geqq \left(\frac{1}{d} - \frac{1}{d^n + d}\right) A_i * A_i \geqq \frac{1}{d+1} A_i * A_i , \qquad 1 < i \leqq r .$$

From Li's equation it follows that

$$E(Z) \geqq p_1 A_1 * A_1$$

$$E(Z) \geqq p_{i-1} A_{i-1} * A_i + p_i A_i * A_i \geqq \left(\frac{1}{d+1} p_{i-1} + p_i\right) A_i * A_i , \qquad 1 < i \leqq r .$$

Dividing the i'th equation by $A_i * A_i = E(T(A_i))$, then summing up one obtains

$$E(Z) \sum_{i=1}^{r} E(T(A_i))^{-1} \geqq 1 + \frac{1}{d+1} - \frac{1}{d+1} p_r . \tag{3}$$

Here

$$p_r \leqq E(Z)/A_r * A_r \leqq \frac{1}{r} E(Z) \sum_{i=1}^{r} E(T(A_i))^{-1} .$$

Substituting this into (3) we arrive at (2):

$$E(Z) \sum_{i=1}^{r} E(T(A_i))^{-1} \geqq \left(1 + \frac{1}{d+1}\right) \Big/ \left(1 + \frac{1}{r(d+1)}\right) \geqq \frac{2d+4}{2d+3} .$$

Since the independence principle fails even in natural cases, it would be desirable to characterize all functionals

$$\varphi_n = \varphi_n(A_1, A_2, \ldots, A_{d^n})$$

to which it is applicable. This problem seems hard and so far I haven't found the answer.

3. Proof of Theorem 1

The aforementioned estimation for the departure of the joint distribution of the waiting times from independence and exponentiality cannot be applied to the whole set H_n, since its accuracy decreases with the increase of the number of words under consideration. It seems better to use the inclusion–exclusion formula, which expresses the distribution function of W_n in terms of the distribution functions of partial minima of the $T(A)$'s. Further, it may be regarded as an expansion into alternating series, which can be well approximated by its sections. But what we in fact need is the graph-sieve of Rényi, which allows us to leave "bad" subsets (with significant overlapping) out of consideration.

For the sake of convenience let us denote $\log|H_n|$ by m. Let the considered words be numbered from 1 to $|H_n|$ and let C_i denote the event $\{T(A_i) > x\}$, $1 \leqq i \leqq |H_n|$, where $x = d^n(m + y)$. Then

$$P(d^{-n} W_n - m \leqq y) = P\left(\bigcap_{i \leqq |H_n|} \bar{C}_i \right) .$$

Let $\varepsilon = \varepsilon_n$ be fixed positive numbers tending to 0 slowly enough:

$$\varepsilon^{-4} m^2 |H_n|^{-\frac{d-1}{d+1}} = o(1) . \tag{4}$$

A word $A \in H_n$ is said to be *bad* if

$$A * A \geqq d^n + \varepsilon^3 m^{-2} d^n ,$$

otherwise good, further, the ordered pair (A, B), where $A, B \in H_n$, $A \neq B$, is said to be *bad* if

$$A * B \geqq \varepsilon m^{-1} d^n .$$

Let the exceptional set E_n defined as

$$E_n = \{(i,j)\colon 1 \leqq i < j \leqq |H_n|, \text{ not all of } A_i, A_j, (A_i, A_j), (A_j, A_i) \text{ are good}\}.$$

Let $S_r^* = \sum_r^* P(C_{i_1} \cap \ldots \cap C_{i_r})$, where $\sum_r^*$ denotes that the summation runs over all r-tuples of indices $1 \leqq i_1 < \ldots < i_r \leqq |H_n|$ which do *not* contain any pairs

from E_n (for $r = 0$ let $S_0^* = 1$). Finally, for $r \geqq 2$ let $S_r^{**} = \sum_r^{**} P(C_{i_1} \cap \ldots \cap C_{i_r})$, where $\sum_r^{**}$ indicates summation over r-tuples containing *exactly one* pair from E_n. Then the graph-sieve of Rényi (see Galambos (1978), Theorem 1.4.2) implies

$$\left| P\left(\bigcap_{i \leqq |H_n|} \bar{C}_i \right) - \sum_{r=0}^{k} (-1)^r S_r^* \right| \leqq S_{k+1}^* + \sum_{r=2}^{k+1} S_r^{**} , \quad k < |H_n| . \tag{5}$$

The proof of Theorem 1 will be carried out by showing that $S_r^* \to \frac{1}{r!} e^{-ry}$ and $S_r^{**} \to 0$ as $n \to \infty$.

Lemma 3. *(a) The number of bad words is less than $m^2\varepsilon^{-4}$. (b) For any given $A \in H_n$ the number of words B for which the pair (A, B) (resp. (B, A)) is bad, is less than $m\varepsilon^{-2}$.*

Proof. (a) Suppose the maximum overlapping of A with itself is of length l (apart from the fact that A is identical with itself which can be interpreted as overlapping of length n), and let k be the minimum of l as A runs over the bad words. Then

$$d^n(1 + \varepsilon^3 m^{-2}) \leqq A * A \leqq d^n + d^k + d^{k-1} + \cdots + d < d^n + 2d^k ,$$

from which $d^{n-k} < 2m^2\varepsilon^{-3}$. The number of words $A \in H^n$ that overlap themselves in length l is d^{n-l}, thus the number of bad words is not greater than

$$d^{n-k} + d^{n-k-1} + \cdots + d < 2d^{n-k} < 4m^2\varepsilon^{-3} < m^2\varepsilon^{-4} .$$

(b) Similarly, let k be the minimum of the longest overlapping between A and B (resp. B and A) as B varies in such a way that (A, B) (resp (B, A)) is bad. Then $\varepsilon m^{-1} d^n < 2d^k$. The number of words $B \in H^n$ that overlap A in length l is d^{n-l} again and the proof can be completed in the same way as above.

Lemma 4. *Let $|\sum_r^{**}|$ and $|\sum_r^{**}|$ denote the number of terms of the corresponding sum. Then*

$$\frac{1}{r!}(|H_n| - rm^2\varepsilon^{-4})^r \leqq |\textstyle\sum_r^*| \leqq \frac{1}{r!}|H_n|^r , \quad r \geqq 1 ,$$

and

$$|\textstyle\sum_r^{**}| \leqq \frac{1}{(r-2)!}|H_n|^{r-1} m\varepsilon^{-2} \quad \text{if } r > 2 .$$

Proof. Estimation of $|\sum_r^*|$. Let us forget the increasing order of indices $i_1, \ldots, i_r$, this will give us a multiplier $r!$. Now the upper bound $|H_n|^r$ is obvious. For the lower bound delete first all the bad words from H_n (by Lemma 3 (a) there are at most $m^2\varepsilon^{-4}$ of them), then choose A_{i_1}. Fixing A_{i_1} one can find at most $2m\varepsilon^{-2}$ bad pairs through it. Thus the number of words that can be chosen as A_{i_2} has decreased by at most $2m\varepsilon^{-2}$ more, hence A_{i_2} is to be chosen from a set of size $\geqq |H_n| - 2m^2\varepsilon^{-4}$. Here we used that $2m\varepsilon^{-2} \leqq m^2\varepsilon^{-4} - 1$ for large enough n. Continuing in this way we see that less than $m^2\varepsilon^{-4}$ words should be deleted at every stage.

Estimation of $|\sum_r^{**}|$. Again, r-tuples in account can not contain any bad words, or else they contained more than one pairs from E_n. By Lemma 3(b) there are at

most $|H_n|m\varepsilon^{-2}$ bad pairs to choose, while for the other $r-2$ indices we have at most $\binom{|H_n|}{r-2}$ choices.

Lemma 5. *Let* $A_1, A_2, \ldots, A_r \in H^n$, $Z = \min T(A_i)$.

(a) *Suppose* $A_i * A_i \leqq (1+\delta)d^n$ *for* $1 \leqq i \leqq r$ *and* $A_i * A_j \leqq \delta d^n$ *for* $i \neq j$. *Then*

$$\frac{1}{r}d^n \leqq E(Z) \leqq \left(\frac{1}{r}+\delta\right)d^n .$$

(b) *Suppose the conditions of* (a) *are met with the only exception* $A_k * A_l = cd^n$, $c > \delta$. *Then*

$$\frac{1-c\delta}{r-c}d^n \leqq E(Z) \leqq \frac{1+r\delta}{r-c}d^n .$$

(c) *Suppose the conditions of* (a) *are met with the only exception* $A_k * A_k = (1+c)d^n$, $c > \delta$. *Then*

$$\left(r-\frac{c}{1+c}\right)^{-1} d^n \leqq E(Z) \leqq \left[\delta + \left(r-\frac{c}{1+c}\right)^{-1}\right] d^n .$$

Proof. These assertions are simple consequences of Lemma 2.

(a) From Li's equations

$$p_j \leqq d^{-n}E(Z) \leqq p_j(1+\delta) + \sum_{i \neq j} p_i\delta = p_j + \delta . \tag{6}$$

Summing up for $j = 1, 2, \ldots, r$ one obtains the desired estimation.

(b) The l'th inequality in (6) must be replaced by

$$p_l + p_k c \leqq d^{-n}E(Z) \leqq p_l + p_k c + \delta .$$

Adding up these inequalities we get

$$1 + p_k c \leqq rd^{-n}E(Z) \leqq 1 + r\delta + p_k c \leqq 1 + r\delta + cd^{-n}E(Z) ,$$

from which the upper bound immediately follows. On the other hand,

$$d^{-n}E(Z) - \delta \leqq p_k ,$$

thus

$$1 + cd^{-n}E(Z) - c\delta \leqq rd^{-n}E(Z) ,$$

hence the lower bound.

(c) For $j = k$ (6) changes into

$$p_k(1+c) \leqq d^{-n}E(Z) \leqq p_k(1+c) + \delta .$$

Dividing it by $1+c$ then summing up we arrive at

$$1 \leqq \left(r-1+\frac{1}{1+c}\right)d^{-n}E(Z) \leqq 1 + \left(r-1+\frac{1}{1+c}\right)\delta ,$$

which was to be proved.

Lemma 6. *For* $r \geqq 1$ $S_r^* \to \frac{1}{r!}e^{-ry}$ *as* $n \to \infty$.

Proof. Let first $r = 1$. From Lemma 1

$$S_1^* = \sum_{A \in H_n} P(T(A) > x) \leqq \frac{1-4b}{1-5b} \sum_{A \in H_n} \exp\left(-\frac{x}{A*A}\right)$$
$$= \frac{1-4b}{1-5b} e^{-y} |H_n|^{-1} \sum_{A \in H_n} \exp\left(d^{-n}x \frac{A*A-d^n}{A*A}\right),$$

where $b = nd^{-n}$.

The sum is divided into two parts accordingly that A is bad or good. Good words give

$$|H_n|^{-1} \sum_{A \text{ good}} \exp\left(d^{-n}x \frac{A*A-d^n}{A*A}\right) \leqq |H_n|^{-1} \sum_{A \text{ good}} \exp(\varepsilon^3 m^{-2} d^{-n}x)$$
$$= 1 + o(1),$$

since $d^{-n}x \sim m$.

The bad words' contribution is

$$|H_n|^{-1} \sum_{A \text{ bad}} \exp\left(d^{-n}x \frac{A*A-d^n}{A*A}\right) \leqq |H_n|^{-1} m^2 \varepsilon^{-4} \exp((m+y)/d)$$
$$= O(m^2 \varepsilon^{-4} |H_n|^{-(d-1)/d}) = o(1).$$

Here we used first Lemma 3, then the fact that $A*A < \frac{d}{d-1} d^n$, finally (4).

Hence

$$S_1^* \leqq e^{-y}(1 + o(1)).$$

On the other hand, Lemma 1 implies that

$$S_1^* = \sum_{A \in H_n} P(T(A) > x) \geqq \sum_{A \in H_n} \exp\left(-\frac{1-4b}{1-5b} \frac{x}{A*A}\right)$$
$$\geqq |H_n| \exp(-(1 + O(b)) d^{-n}x) = e^{-y}(1 + o(1)).$$

Let us turn to the case $r \geqq 2$. Putting $b = rnd^{-n}$, $\delta = \varepsilon/m$ and combining Lemma 1 with Lemmas 4 and 5(a) we have

$$S_r^* \leqq \frac{1-4b}{1-5b} \frac{1}{r!} |H_n|^r \exp(-rd^{-n}x(1 + \varepsilon/m)^{-1})$$
$$= (1 + o(1)) \frac{1}{r!} |H_n|^r \exp(-r(m+y)(1 + o(1/m)))$$
$$= (1 + o(1)) \frac{1}{r!} e^{-ry},$$

and at the same time

$$S_r^* \geqq \frac{1}{r!} (|H_n| - rm^2 \varepsilon^{-4})^r \exp\left(-\frac{1-4b}{1-5b} rd^{-n}x\right)$$
$$= (1 + o(1)) \frac{1}{r!} |H_n|^r \exp(-r(m+y)(1 + o(1/m)))$$
$$= (1 + o(1)) \frac{1}{r!} e^{-ry}.$$

Lemma 7. *For $r \geqq 2$ $S_r^{**} \to 0$ as $n \to \infty$.*

Proof. Let first $r > 2$. Any r-tuple of indices taken into summation contains exactly one pair from E_n, say (i, j). Then both A_i and A_j have to be good, further, (A_i, A_j) and (A_j, A_i) cannot be bad at the same time, since if both pairs were bad, i.e.

$$A_i * A_j \geqq d^n \varepsilon/m \quad \text{and} \quad A_j * A_i \geqq d^n \varepsilon/m ,$$

then A_i would overlap A_j and conversely, at least in length k, where $\frac{d}{d-1} d^k > d^n \varepsilon/m$. But this would imply that A_i overlaps itself in length not less than $2k - n$, and consequently,

$$A_i * A_i - d^n \geqq d^{2k-n} > \left(\frac{d-1}{d}\right)^2 \varepsilon^2 m^{-2} d^n > \varepsilon^3 m^{-2} d^n ,$$

contradicting the goodness of A_i. Now let us apply Lemmas 1, 4 and 5(b) with $\delta = \varepsilon/m$. We shall separately treat the summands whose bad pair (A_i, A_j) is of maximal leading number $d^{n-1} + d^{n-2} + \cdots + d$. There are d^3 of them, because in these pairs the same letter has to stand at every position but the first in A_i and the last in A_j. All the other pairs have leading number $A_i * A_j \leqq d^{n-1} + d^{n-3} + d^{n-5} + \cdots < \frac{d}{d^2-1} d^n$. Hence

$$S_r^{**} \leqq \frac{1}{(r-2)!} d^3 |H_n|^{r-2} \frac{1-4b}{1-5b} \exp\left(-d^{-n}x\left(r - \frac{1}{d-1}\right)(1 + r\varepsilon/m)^{-1}\right)$$

$$+ \frac{1}{(r-2)!} m\varepsilon^{-2} |H_n|^{r-1} \frac{1-4b}{1-5b} \exp\left(-d^{-n}x\left(r - \frac{d}{d^2-1}\right)(1 + r\varepsilon/m)^{-1}\right).$$

Since $d^{-n}x \sim m$, the first term can be transformed into

$$O(|H_n|^{-2+\frac{1}{d-1}}) ,$$

while the second one into

$$O(|H_n|^{-1+\frac{1}{d^2-1}} m\varepsilon^{-2}) .$$

Both expressions tend to 0 by (4).

In the case $r = 2$ we can't come to the conclusion that both words should be good. For this reason the sum $\sum_2^{**}$ will be divided into three parts. Summands, where both words are good, can be treated in the same way as for $r > 2$. If one of the words is good, one is bad and the pair they form is good, whatever be their order, then Lemma 5(c) with $c < \frac{1}{d-1}$, $\delta = \varepsilon/m$ gives

$$E(Z) \leqq d^n \left(\frac{d}{2d-1} + \frac{\varepsilon}{m}\right).$$

The number of such summands is less than $m^2\varepsilon^{-4}|H_n|$. By Lemma 1 this part of $\sum_2^{**}$ is bounded by

$$m^2\varepsilon^{-4}|H_n| \frac{1-4b}{1-5b} \exp\left(-d^{-n}x\left(\frac{d}{2d-1} + \frac{\varepsilon}{m}\right)^{-1}\right) = O(|H_n|^{-(d-1)/d} m^2\varepsilon^{-4}) = o(1) .$$

(We applied (4) again.)

Finally, there are summands where either both words are bad or one of them is bad and the pair is also bad in some order. The number of these summands is bounded by

$$(m^2\varepsilon^{-4})^2 + 2m^2\varepsilon^{-4}m\varepsilon^{-2} < 3m^4\varepsilon^{-8} .$$

For such pairs Lemma 5(a) with $\delta = \dfrac{1}{d-1}$ gives

$$E(Z) \leqq \frac{d+1}{2(d-1)} d^n ,$$

hence their contribution is less than

$$3m^4\varepsilon^{-8} \frac{1-4b}{1-5b} \exp\left(-2d^{-n}x \frac{d-1}{d+1}\right) = O([|H_n|^{-\frac{d-1}{d+1}} m^2\varepsilon^{-4}]^2) = o(1) .$$

Thus the proof of Lemma 7 is completed.

Now we are in a position to complete the proof of Theorem 1. Fixing k and applying Lemmas 6 and 7 to the terms of (5) we obtain that

$$\sum_{r=0}^{k} (-1)^r \frac{1}{r!} e^{-ry} - \frac{1}{(k+1)!} e^{-(k+1)y} \leqq \liminf_{n\to\infty} P(d^{-n} W_n - m \leqq y)$$

$$\leqq \limsup_{n\to\infty} P(d^{-n} W_n - m \leqq y) \leqq \sum_{r=0}^{k} (-1)^r \frac{1}{r!} e^{-ry} + \frac{1}{(k+1)!} e^{-(k+1)y} .$$

Letting k tend to infinity we can see that

$$\lim_{n\to\infty} P(d^{-n} W_n - m \leqq y) = \sum_{r=0}^{\infty} (-1)^r \frac{1}{r!} e^{-ry} ,$$

which was to be proved.

4. Further results, remarks and problems

Theorem 1 can be generalized in several ways. For example, let $W_n(k)$ denote the waiting time till all but k words of H_n have been observed as a run. Clearly, $W_n = W_n(0)$. Then we have the following assertion.

Theorem 2. *Suppose*

$$\lim_{n\to\infty} |H_n| = +\infty .$$

Then for $k = 0, 1, 2, \ldots$

$$\lim_{n\to\infty} P(d^{-n} W_n(k) - \log|H_n| \leqq y) = e^{-e^{-y}} \sum_{r=0}^{k} \frac{1}{r!} e^{-ry} .$$

The proof of Theorem 2 is similar to that of Theorem 1. The only change is that instead of the graph-sieve of Rényi we have to apply its extension by Galambos (1966) and the analogue of Lemma 7 needs a bit more technique and patience (increasing with k).

Another way of extension is estimating the rate of convergence, taking also large deviations into consideration, then analysing the a.s. behaviour of the

sequence W_n. Our proof in Sect. 3, refined to a certain extent, is suitable for the above programme. I am planning to return to these questions in a forthcoming paper.

Through the rest of the paper let $H_n = H^n$.

A quantity related to W_n is M_n, the maximal number that every word of length M_n was observed as a run in course of the first n experiments. An immediate consequence of Theorem 1 is the following assertion.

Theorem 3. *Let* $k = \left[\dfrac{\log n - \log\log n}{\log d}\right]$, *where* $[.]$ *stands for integer part. Then*

$$\lim_{n\to\infty} P(M_n = k \text{ or } k+1) = 1 .$$

Proof. For every real y and for k and n large enough, depending one on another as above we have

$$d^k(\log d^k + y) < n < d^{k+2}(\log d^{k+2} + y) .$$

In order to verify this let us rewrite the above definition of k into

$$\log d^k \leqq \log n - \log\log n < \log d^{k+1} .$$

Since $\log n - \log\log n$ is increasing in n, it suffices to show that putting $n = d^k(\log d^k + y)$ we have $\log n - \log\log n < \log d^k$, and conversely, $n = d^{k+2}(\log d^{k+2} + y)$ implies $\log n - \log\log n \geqq \log d^{k+1}$. Details of this straightforward calculation are left to the reader.

The random variables W_k and M_n are connected by the following relation

$$W_k \leqq n \quad \text{iff} \quad M_n \geqq k .$$

Hence by Theorem 1

$$P(M_n \geqq k) = P(W_k \leqq n) \geqq P(d^{-k}W_k - \log d^k \leqq y) \to e^{-e^{-y}} ,$$

and

$$P(M_n \geqq k+2) = P(W_{k+2} \leqq n) \leqq P(d^{-k+2}W_{k+2} - \log d^{k+2} \leqq y) \to e^{-e^{-y}} .$$

Since y is arbitrary, the proof is completed.

We should remark that Theorem 3 can be strengthened. It can be proved that $M_n = \left[\dfrac{\log n - \log\log n - \varepsilon}{\log d}\right]$ or $M_n = \left[\dfrac{\log n - \log\log n + \varepsilon}{\log d}\right]$ for large n with probability one.

Another related problem is to find sharp bounds for $E(W_n)$. If in Theorem 1 the variables $d^{-n}W_n - \log d^n$ were uniformly integrable,

$$E(W_n) = d^n(\log d^n + C + o(1))$$

would follow, where $C = 0.577\ldots$ the Euler-Mascheroni constant. By another, direct method, in Móri (1987) it is proved that

$$E(W_n) = d^n(\log d^n + O(1)) .$$

Finally, we set a problem, which probably requires new ideas in addition to those of the present paper.

What can be said about W_n if $X_1, X_2, \ldots$ are identically but *not uniformly* distributed on H? In this case the $T(A)$ are still exponentially distributed in the

limit, but no more with the same expectation, which probably kills the graph-sieve. What remains true is the Conway algorithm and the asymptotic exponentiality.

Acknowledgement. Research supported by the Hungarian National Foundation for Scientific Research, Grant No. 1808.

References

Erdös, P., Rényi, A.: On a new law of large numbers. J. Anal. Math. **23**, 103–111 (1970)
Erdös, P., Révész, P.: On the length of the longest head-run. In: Coll. Math. Soc. J. Bolyai 16, Topics in Information Theory, Keszthely, Hungary 1975. Amsterdam: North Holland 1976
Galambos, J.: On the sieve methods in probability theory I. Stud. Sci. Math. Hung. **1**, 39–50 (1966)
Galambos, J.: The asymptotic theory of extreme order statistics. New York: Wiley 1978
Guibas, L.J., Odlyzko, A.M.: String overlaps, pattern matching and nontransitive games. J. Comb. Theory, Ser. A **30**, 183–208 (1981)
Li, S.R.: A martingale approach to the study of occurrence of sequence patterns in repeated experiments. Ann. Probab. **8**, 1171–1176 (1980)
Móri, T.F.: Large deviation results for waiting times in repeated experiments. Acta Math. Acad. Sci. Hungar. **45**, 213–221 (1985)
Móri, T.F.: On the expectation of the maximum waiting time. Ann. Univ. Sci. Budap. Rolando Eötvös, Sect. Comput. **7**, 111–115 (1987)
Móri, T.F., Székely, G.J.: Asymptotic independence of 'pure head' stopping times. Stat. Probab. Lett. **2**, 5–8 (1984)
Shukle, R., Shrivastava, R.C.: The statistical analysis of direct repeats in nucleic acid sequences. J. Appl. Probab. **22**, 15–24 (1985)