Probab. Th. Rel. Fields 72, 279–287 (1986)

Probability Theory and Related Fields

An Extreme Value Theory for Long Head Runs*

Louis Gordon, Mark F. Schilling, and Michael S. Waterman

Department of Mathematics, University of Southern California,
Los Angeles, California 90089-1113, USA

Summary. For an infinite sequence of independent coin tosses with $P(\text{Heads})=p\in(0,1)$, the longest run of consecutive heads in the first n tosses is a natural object of study. We show that the probabilistic behavior of the length of the longest pure head run is closely approximated by that of the greatest integer function of the maximum of $n(1-p)$ i.i.d. exponential random variables. These results are extended to the case of the longest head run interrupted by k tails. The mean length of this run is shown to be $\log(n)+k\log\log(n)+(k+1)\log(1-p)-\log(k!)+k+\gamma/\lambda-1/2+r_1(n)+o(1)$ where $\log=\log_{1/p}$, $\gamma=0.577\ldots$ is the Euler-Mascheroni constant, $\lambda=\ln(1/p)$, and $r_1(n)$ is small. The variance is $\pi^2/6\lambda^2+1/12+r_2(n)+o(1)$, where $r_2(n)$ is again small. Upper and lower class results for these run lengths are also obtained and extensions discussed.

1. Introduction

Consider an infinite sequence of independent coin tosses of a possibly biased coin, in which heads appear with probability p and tails appear with probability q. The longest run of consecutive heads in the first n tosses is a natural object of study. Several strong results for this and related problems have been obtained only rather recently. Erdös and Révész [6] employ counting arguments to provide almost sure results about the growth of the longest head run when $p=q=1/2$. Among the curiosities that such runs display is that the upper class boundary for unexpectedly long longest head runs grows like $\ln(n)/\lambda+\ln\ln(n)/\lambda$, while the lower class boundary for unusually short longest head runs grows like $\ln(n)/\lambda-\ln\ln\ln(n)/\lambda$, where $\lambda=\ln(1/p)$. Erdös and Révész also extend their results to the case of the longest k-interrupted head run, in which up to k tails may appear among a run of heads.

* This work was supported by a grant from the System Development Foundation

Guibas and Odlyzko [8] extend and deepen these results to include arbitrary repetitive patterns of heads and tails. By an impressive use of transform techniques, they explicitly compute very tight approximations to the mean and variance of the longest pure head run. Denote the longest pure head run in the first n tosses by $Z_0(n)$. Guibas and Odlyzko show that in the case $p=q=1/2$, $EZ_0(n)=(\ln(n)+\gamma)/\lambda-3/2+\rho_0(n)+o(1)$ where $\gamma=0.577\ldots$ is the Euler-Mascheroni constant and $\lambda=\ln(2)$. The quantity $\rho_0(n)$, although small ($|\rho_0(n)|<1.6\times 10^{-6}$ when $p=1/2$), possesses no limit. This leads to the intriguing conclusion that the longest head run possesses no limit distribution. An attempt to give this phenomenon a probabilistically intuitive explanation was a large motivation of our work.

The key observation in our approach is that the length of any single pure head run is distributed as a geometric random variable. Denote by $N(n)$ the number of head runs in the first n tosses. The longest head run is then the maximum of a random number $N(n)$ of independent geometric random variables. Furthermore, $N(n)$ obeys various strong laws of large numbers; indeed it is binomial (n, q). In addition, the geometric distribution with parameter q can be represented as the integer part of an exponential random variable with mean $1/\lambda$, where $\lambda=\ln(1/p)$. Anderson [1] uses this relation to investigate the maxima of independent integer-valued random variables. Anderson explicitly studies the maxima of independent geometric random variables, a situation for which there is no limiting distribution. Ferguson [7] provides additional limit theory for approximating the distribution of maxima of integer-valued random variables with exponentially decaying tails.

We show below that the behavior of the longest pure head run $Z_0(n)$ is very closely approximated by that of $\lfloor \max_{j\leqq nq} \hat{Y}_0(j) \rfloor$ where $\hat{Y}_0(j)$ are i.i.d. exponential random variables with mean $1/\lambda$ and $\lfloor\cdot\rfloor$ is the greatest integer function. Letting W denote a random variable with the standard extreme value distribution, i.e. $P\{W\leqq t\}=\exp(-e^{-t})$, we can then approximate the distribution of $\max_{j\leqq nq} \hat{Y}_0(j)$ by that of $W/\lambda+\ln(nq)/\lambda$. Hence $EZ_0(n)$ is almost $EW/\lambda+\ln(nq)/\lambda-1/2$, where $EW=\gamma$, and the term $1/2$ is Sheppard's correction for continuity (Kendall and Stuart ([9], p. 77)). In the case $p=q=1/2$ the approximation becomes $(\gamma+\ln(n))/\lambda-3/2$, precisely the leading terms of Guibas and Odlyzko. Similarly, applying to W/λ Sheppard's correction for the variance yields an approximate variance of $\pi^2/6\lambda^2+1/12$. This approximation, with much smaller higher order terms, is again found in Guibas and Odlyzko.

Our program is as follows. In Sect. 2, we formalize a representation of coin tossing and establish the notation we use throughout. We also generalize to the negative binomial case Anderson's representation of geometric random variables as integerized exponential random variables. This generalization lets us handle k-interrupted head runs. Approximation in distribution is the goal of Sect. 3. The key tool is the result of Watson [15] on the maximum of k-dependent random variables. This form of dependence is introduced because we study k-interrupted head runs, which are essentially moving sums of $k+1$ adjacent i.i.d. geometric random variables.

In Sects. 4 and 5 we obtain upper and lower class results for k-interrupted head runs. These sections are anticipated by Deheuvel's [5] comprehensive

study of multidimensional Erdös-Renyi theorems. Formulating the problem in terms of maxima of independent or k-dependent random variables allows us to use work of Robbins and Siegmund [11] on the law of the iterated logarithm for maxima. O'Brien [10] generalizes their work to dependent variables under strong mixing conditions. As does O'Brien, we remark that Barndorff-Nielsen's [2] theorems are similar to Robbins and Siegmund's, though the Robbins and Siegmund hypotheses are more convenient for our purposes. Sections 4 and 5, especially the statement of our Theorem 4, owe much to O'Brien's work.

In Sect. 6, we conclude by discussing how our results could be generalized to situations of Markov dependence and to behavior of repetitive patterns as in Solov'ev [14].

2. The Fundamental Representation

Let $Y_0(1), Y_0(2), \ldots$ be an i.i.d. sequence of geometric random variables with parameter q, so that $P\{Y_0(n)=m\}=qp^m$, for m a non-negative integer and $p=1-q$. Write $S(m)=m+\sum_{j=1}^{m} Y_0(j)$ for $m\geqq 1$ and $S(m)=0$ for $m\leqq 0$. We now realize independent Bernoulli (p) random variables as $X_n=I_{\{n=S(j) \text{ for all } j>0\}}$.

The values $\{S(m)\}$, for $m>0$, are the locations of T's (tails when $X_n=0$) in the sequence $X_1X_2X_3\ldots$ of H's (heads when $X_n=1$) and T's. Write $N(n)$ for the binomial (n,q) number of T's observed in the first n tosses. $Y_0(m)$ is the length of the m-th completed pure head run, which is ended by a T at the $S(m)$-th toss. Note that runs of length 0 are permitted and that the length of the last head run in n tosses is $n-S(N(n))\leqq Y_0(N(n)+1)$.

We now consider head runs interrupted by kT's. For $m>0$, we write $Y_k(m)=S(m+k)-(1+S(m-1))$ for the summed lengths of the $k+1$ pure head runs starting with run m and ending with run $m+k$, plus the k intervening T's which separate the $k+1$ component pure head runs. Note that $Y_k(m)-k$ has a negative binomial distribution with parameters $(k+1,q)$. We set $Y_k(m)=0$ for $m\leqq 0$.

Write $M_k(n)=\max_{m\leqq n} Y_k(m)$ when $n>0$, and set $M_k(n)=0$ for $n\leqq 0$. The length of the longest k-interrupted head run in the first n tosses is then denoted $Z_k(n)=\max\{M_k(N(n)-k), n-S(N(n)-k)\}$. Note that our $Z_k(n)$ corresponds to the $Z_n(k)$ of Erdös and Révész.

It is clear from the construction that the behavior of the longest k-interrupted head run is strongly related to the behavior of the maxima of $k+1$-moving sums of independent geometric random variables. We exploit this relationship in subsequent sections.

3. Approximation in Distribution

Our goal in this section is to obtain results concerning the approximate distribution of the longest k-interrupted head run $Z_k(n)$. The preceding repre-

sentation reduces the problem to a study of the extreme tails of the negative binomial distribution.

Let $\lambda = \ln(1/p)$ and write $\log(n)$ for $\log_{1/p}(n)$. Define the constants $\mu_k(n) = \log(n) + k\log\log(n) + k\log(q/p) - \log(k!)$. We construct below a stationary k-dependent sequence of continuously distributed random variables $\{\hat{Y}_k(j)\}$ such that $\lfloor \hat{Y}_k(j) \rfloor = Y_k(j) - k$ and $\max_{j \leqq n} \{\hat{Y}_k(j)\} + k - \mu(qn) \to W/\lambda$ in distribution, where W has a standard extreme value distribution.

It is therefore plausible that $Z_k(n)$ is well approximated by $M_k(N(n))$. The law of large numbers tells us that $N(n)$ is close to qn. Hence we expect that $Z_k(n)$ has a distribution close to that of $\lfloor W/\lambda + \mu_k(qn) \rfloor$. We formalize this reasoning in Theorem 1.

In Theorem 2, we emphasize the probabilistic content implicit in the calculations of Boyd [3] and of Guibas and Odlyzko [8]. By examining the remainder term in Sheppard's correction, we show that $E\{Z_k(n)\}$ is approximately $\mu_k(qn) + \gamma - 1/2$, and that $\mathrm{Var}\{Z_k(n)\}$ is about $\pi^2/(6\lambda^2) + 1/12$. Note that $E\{W\} = \gamma$ and $\mathrm{Var}\{W\} = \pi^2/6$.

Our construction depends on the following observations, collected in Lemma 1, below. Note that for integer x, $P\{Y_k(1) - k \geqq x\} = (k!)^{-1} q^{k+1} (d/dp)^k (p^{x+k}/(1-p)) = Q_k(x) p^x$, where $Q_k(x)$ is a k-degree polynomial whose leading term is $(qx)^k/k!$. Hence for all x sufficiently large and for all integer $x \geqq 0$, there exists a strictly decreasing continuous function $G_k(x) = (k!)^{-1} q^{k+1} (d/dp)^k (p^{x+k}/(1-p)) = Q_k(x) p^x$.

Hence, we study the tail behavior of the negative binomial by examining the properties of continuous random variables with distribution $1 - G_k(\cdot)$. These properties are summarized in Lemma 1.

Lemma 1. a) $G_k(x)((qx)^k p^x/k!)^{-1} = 1 + O(1/x)$.

b) $nG_k(\zeta + \mu_k(n) - k) \to p^\zeta$ *as* $n \to \infty$.

c) *If* $\hat{Y}$ *has distribution* $1 - G_k(\cdot)$, *then* $\lfloor \hat{Y} \rfloor$ *has a negative binomial* $(k+1, q)$ *distribution.*

By construction as in O'Brien [10], we may work in a probability space for which $Y_k(n)$, X_m and $Z_k(n)$ are defined as in Sect. 2, and for which there is a k-dependent sequence $\hat{Y}_k(n)$ sharing the continuous marginal distribution $1 - G_k(\cdot)$. Further, $\lfloor \hat{Y}_k(n) \rfloor + k = Y_k(n)$. For the case $k = 0$, corresponding to pure head runs, note that $\hat{Y}_k(n)$ are i.i.d. exponential random variables with mean $1/\lambda$.

By analogy with Sect. 2, we write $\hat{M}_k(n) = \max_{j \leqq n} \{\hat{Y}_k(j) + k\}$. We first use the construction to approximate the longest k-interrupted head run.

Theorem 1. *Let* $\rho_k(n) = \mu_k(n) - \lfloor \mu_k(n) \rfloor$ *and let* W *have a standard extreme value distribution. Then, uniformly in* t,

$$P\{Z_k(n) - \mu_k(qn) \leqq t\} - P\{\lfloor W/\lambda + \rho_k(qn) \rfloor - \rho_k(qn) \leqq t\} \to 0$$

as $n \to \infty$.

Proof. We verify the mixing condition of Watson [15]. Let $1 \leqq l \leqq k$ and let U_1, U_2, U_3 be independent negative binomial random variables with respective

parameters (l,q), $(k+1-l,q)$ and (l,q). We show that $P\{U_1+U_2\geqq m$ and $U_2+U_3\geqq m|U_1+U_2\geqq m\}\to 0$ as $m\to 0$ along the integers. Let $\tau=\lfloor\log(m)/2\rfloor$. From Lemma 1, $P\{U_1+U_2\geqq m\}\geqq C_1 m^k p^m$. In addition, $P\{U_1+U_2\geqq m$ and $U_2+U_3\geqq m\}$

$$\leqq P\{U_2\geqq m-\tau\}+\sum_{n=0}^{m-\tau} P^2\{U_1\geqq m-n\}P\{U_2=n\}\leqq C_2 m^{-1/2}m^k p^m,$$

where $C_i=C_i(n,p)$ denote constants. Hence Watson's result is applicable to $\hat{M}_k(n)=k+\max_{j\leqq n}\{\hat{Y}_k(j)\}$.

Write $\delta(n)=n^{1/2}\log^2(n)$. Because $N(n)$ is binomial, $P\{|N(n)-nq|>\delta(n)-1\}\to 0$. Hence, with arbitrarily large probability, $\lfloor\hat{M}_k(\lfloor nq-\delta(n)\rfloor)\rfloor\leqq Z_k(n)\leqq\lfloor\hat{M}_k(\lfloor nq+\delta(n)\rfloor)\rfloor$.

We use k-dependence and argue along subsequences of independent (and so exchangeable) random variables to obtain:

$$P\{\hat{M}_k(\lfloor nq-\delta(n)\rfloor)=\hat{M}_k(\lfloor nq+\delta(n)\rfloor)\}\to 1.$$

Now Watson's result and Lemma 1 yields $\hat{M}_k(nq)-\mu_k(nq)\to W/\lambda$ in distribution. Because W has a continuous distribution, $P\{\lfloor\hat{M}_k(nq)\rfloor-\lfloor\mu_k(nq)\rfloor\leqq t\}-P\{\lfloor W/\lambda+\rho_k(nq)\rfloor\leqq t\}\to 0$ uniformly in t. The theorem follows because $Z_k(n)$ equals $\lfloor\hat{M}_k(nq)\rfloor$ with probability approaching 1.

Note that we have not proved that $Z_k(n)$ has a limiting distribution. That the failure is not severe is the import of Theorem 2, in which we approximate the mean and variance of $Z_k(n)$. This calculation is a probabilistic version of the generating function calculations of Boyd and of Guibas and Odlyzko. In effect, we obtain explicit bounds for the remainder in Sheppard's correction. (See e.g., Kendall and Stuart [9], p. 77.)

Theorem 2. *Let* $\theta=\pi^2/\lambda$. *Then*

$$|E\{Z_k(n)\}-(\mu_k(qn)+\gamma/\lambda-1/2)|<(2\pi)^{-1}\theta^{1/2}e^{-\theta}(1-e^{-\theta})^{-2}+o(1)$$

and

$$|\mathrm{Var}\{Z_k(n)\}-(\pi^2/(6\lambda^2)+1/12)|<2(1.1+0.7\theta)\theta^{1/2}e^{-\theta}(1-e^{-\theta})^{-3}+o(1).$$

Proof. As a preliminary, we establish the uniform integrability of the sequence $\{Z_k(n)-\mu_k(nq)\}$. Observe that, for given $t>0$,

$$P\{Z_k(n)-\mu_k(nq)>t\}\leqq nP\{\hat{Y}_k(1)-\mu_k(nq)>t\}\leqq C_1(1+t)^k p^t,$$

and that

$$\begin{aligned}P\{Z_k(n)-\mu_k(nq)<-t\}&\leqq\exp(-C_2 nG_k(\mu_k(nq)-t))+P\{N(n)<nq/2\}\\&\leqq\exp(-C_3 p^{-t})+\exp(-nq/8)\end{aligned}$$

for positive constants C_1, C_2, and C_3. Also, $Z_k(n)\leqq n$. Hence $\{Z_k(n)-\mu_k(nq)\}$ and its squares are uniformly integrable.

Now consider the two Bernoulli functions, $b_1(x)=x-\lfloor x\rfloor-1/2$ and $b_2(x)=b_1^2(x)-1/12$. Because of Theorem 1 and uniform integrability, we may evalu-

ate $E\{b_1(W/\lambda+\rho)\}$, $\mathrm{Var}\{b_1(W/\lambda+\rho)\}$, and $\mathrm{Cov}\{W, b_1(W/\lambda+\rho)\}$ for constants λ and ρ.

Expand b_1 and b_2 in Fourier series (Rogosinski [12], pp. 38, 135) and apply dominated convergence to express the expectations in terms of the characteristic function of W, which is $\Gamma(1-it/\lambda)$. The inequality follows from the reflection formula for the gamma function, and by the series development of the digamma function (Carrier et al. [4], pp. 187, 189).

In the case of fair coin tossing, the bounds are on the order of 1.6×10^{-6} for the mean and 6×10^{-5} for the variance. See Boyd [3], p. 15, for an analytic derivation of very similar bounds.

4. Unusually Long Longest Head Runs

In this section, we begin to study the almost sure behavior of k-interrupted longest head runs. We present in Theorem 3 a complete characterization of the functions which are touched infinitely often (i.o.) by unusually long longest head runs. Theorem 3 is proved by Erdös and Révész for $p=1/2$. See their Theorem 3* and Theorem 4*. Guibas and Odlyzko study purely repetitive runs and give a related result by generating function methods. Our proof is a direct application of the work of Robbins and Siegmund [11]. See also Deheuvels [5].

Theorem 3. *Let $\{d(n)\}$ be a non-decreasing sequence of integers. Then $P\{Z_k(n)\geqq d(n)\text{ i.o.}\}$ is 0 or 1 as $\sum d^k(n)p^{d(n)}$ is finite or infinite.*

Proof. As in Robbins and Siegmund, note that $P\{\hat{M}_k(n)\geqq d(n)\text{ i.o.}\} = P\{\hat{Y}_k(n)\geqq d(n)-k\text{ i.o.}\}$ because $\{d(n)\}$ is non-decreasing. Recall from Sect. 3 that the $\hat{Y}_k(n)$ are k-dependent with common distribution function $1-G_k(\cdot)$.

If $\Sigma G_k(d(n)-k)=\infty$, apply the strong law of large numbers to the binomial sequence $N(n)$ to conclude that

$$P\{Z_k(n)\geqq d(n)\text{ i.o.}\}\geqq P\{\hat{M}_k(n)\geqq d(\lfloor 2n/q\rfloor)\text{ i.o.}\}$$
$$\geqq P\{\hat{Y}_k(n(k+1))\geqq d(\lfloor 2n(k+1)/q\rfloor)-k\text{ i.o.}\}.$$

The latter probability is 1 by virtue of the Borel-Cantelli lemma for independent events and the monotonicity of $\{d(n)\}$.

Similarly, if $\Sigma G_k(d(n)-k)$ is finite,

$$P\{Z_k(n)\geqq d(n)\text{ i.o.}\}\leqq P\{\hat{M}_k(n)\geqq d(n)\text{ i.o.}\}$$
$$\leqq P\{\hat{Y}_k(n)\geqq d(n)-k\text{ i.o.}\}=0.$$

From Lemma 1, $\Sigma G_k(d(n)-k)=\infty$ if and only if $\Sigma d(n)^k p^{d(n)}=\infty$.

5. Unusually Short Longest Head Runs

We now study the almost sure behavior of unusually short longest k-interrupted head runs. From the introduction, recall the asymmetry in the

behavior of unusually short and unusually long longest head runs, as seen in Erdös and Révész's Theorem 2*.

In Theorem 4 we show that $P\{Z_k(n)<d(n) \text{ i.o.}\}$ is 0 or 1 as determined for integer sequences $\{d(n)\}$ by the finiteness of the series $\Sigma d(n)^k p^{d(n)} \exp(-nq(qd(n)/p)^k p^{d(n)}/k!)$. There are a few regularity conditions on $\{d(n)\}$ which we do not mention yet.

It is instructive to compare the series criterion to Erdös and Révész's Theorems 1* and 2*. They study the integer sequences

$$\lfloor \{\log(n)+k\log\log(n)-\log(k!)-1\}-\{\log\ln\log(n)+\delta\}\rfloor.$$

For $\varepsilon>0$, they show that $Z_k(n)$ is strictly below this integer sequence infinitely often when $\delta=1+\varepsilon$, and finitely often when $\delta=-\varepsilon$. The first term is asymptotically equivalent to our $\mu_k(qn)$, when $p=1/2$. Using $x-1<\lfloor x\rfloor\leqq x$, the second term acts like $-\log\ln\log(n)\pm\varepsilon$, where the -1 is used to control the effect of the greatest integer function. The finiteness of the criterion series is therefore controlled by the series $\Sigma n^{-1}\exp(-\ln\log(n)p^{\pm\varepsilon})$. In the practical sense that the small degree of indeterminacy in Erdös and Révész's bounds is due to their explicit handling of the greatest integer function, their result is best possible.

A result like Theorem 4 for purely repetitive patterns and for $p=1/2$ appears in Guibas and Odlyzko. Their series criterion is formulated in terms of a sequence inverse to $\{d(n)\}$.

The key to our approach is the representation of Sect. 3 and Robbins and Siegmund's Theorem 1. ii), of which we use the hypothesis (2.6) in a slightly stronger form. The stronger form is due to O'Brien [10]; for the independent case, see his Theorem 5. We state without proof the slight variant we need as Lemma 2.

Lemma 2. *Let $U_1, U_2, \ldots$ be a stationary k-dependent sequence of random variables with continuous distribution function $1-H(\cdot)$. Let $d(1), d(2), \ldots$ be a non-decreasing sequence of constants such that $\liminf nH(d(n))/\ln\ln(n)>0$. Then $P\{\max_{j\leqq n} U_j<d(n) \text{ i.o.}\}$ is 0 or 1, according to the finiteness of $\Sigma H(d(n))e^{-nH(d(n))}$.*

The precise conditions characterizing the almost sure behavior of unusually short longest head runs are now given in Theorem 4.

Theorem 4. *Let $d(n)$ be a non-decreasing sequence of integers with $\liminf nG_k(d(n))/\log\log(n)>0$ and $\limsup nG_k(d(n))/d(n)<\infty$. Then $P\{Z_k(n)<d(n) \text{ i.o.}\}$ is 0 or 1, as determined by whether $\Sigma d(n)^k p^{d(n)} \cdot \exp(-nq(qd(n)/p)^k p^{d(n)}/k!)$ is finite or infinite.*

Proof. The Markov inequality applied to $m(t)=E(e^{tN(n)})$ implies that $P\{|N(n)-nq|>\delta(n)-1 \text{ i.o.}\}=0$, where $\delta(n)=n^{1/2}\log^2(n)$. Extend $\hat{M}_k(\cdot)$ to the reals by writing $\hat{M}_k(t)=\hat{M}_k(\lfloor t\rfloor)$. Hence,

$$\begin{aligned} P\{\hat{M}_k(nq+\delta(n))<d(n) \text{ i.o.}\} &\leqq P\{Z_k(n)<d(n) \text{ i.o.}\} \\ &\leqq P\{\hat{M}_k(nq-\delta(n))<d(n) \text{ i.o.}\}. \end{aligned}$$

Assume now that the criterion sum is finite. Choose $m(j)$ to be the smallest solution to $m(j)=\lfloor jq-\delta(j)\rfloor$, for j sufficiently large. Note that $m(j)$ is non-decreasing and so

$$P\{\hat{M}_k(nq-\delta(n))<d(n) \text{ i.o.}\}=P\{\hat{M}_k(j)<d(m(j)) \text{ i.o.}\}.$$

From Lemma 1 and the hypothesis, $\liminf G_k(d(m(j))-k)/\log\log(j)$ is positive because $m(j)>jq/2$ for all j sufficiently large.

From Lemma 2, $P\{\hat{M}_k(nq-\delta(n))<d(n) \text{ i.o.}\}=0$ if the sum

$$\Sigma G_k(d(m(j))-k)\exp(-jG_k(d(m(j))-k)$$
$$<e\,\Sigma G_k(d(n)-k)\exp[-(nq+\delta(n))G_k(d(n)-k)]<\infty.$$

However, the latter sum is bounded above by the criterion sum of the hypothesis.

Similarly, if we assume the criterion sum is infinite, Lemma 2 and the identical argument along subsequences establishes the remaining assertion of the theorem.

6. Extensions

In this section, we briefly indicate possible generalizations of the methods we have developed. In the case of finite alphabets with independently chosen letters, one can easily extend the results to purely repetitive patterns of a single word, or of several words which are permutations of the same letters. The analysis then involves finding $N(n)$, the number of starts for the repetitive pattern of interest, and finding a sequence of independent or k-dependent geometric random variables which continue the base word. The limiting behavior is again integerized extreme value.

For example, consider the base word $HHTT$, in a two-letter alphabet. One would expect np^2q^2 occurrences of the base word in the first n tosses. Of these, approximately $n(1-p^2q^2)p^2q^2$ begin a run of the base word. The number of pure repetitions of the base word is thus geometric, and the $N(n)$ such geometrics are independent, because a run must end before a new run begins. Hence, the distribution of the number of words in the longest complete repetitive patterns is approximately distributed as $\lfloor \ln(n(1-p^2q^2))/\lambda+W/\lambda\rfloor$, where $\lambda=-\ln(p^2q^2)$ and W is as before.

In the symmetric case, where all letters of the alphabet are equally likely, one can deal with repetitive patterns with partial repeats adjoined at the ends, as in the paper of Guibas and Odlyzko. In this case, the occurrence of a pattern establishes a template which is followed for a geometrically distributed length.

For example, consider a symmetrically distributed two-letter alphabet $\{H,T\}$, and base word $B=HHTT$. The length in letters of the longest B run (see Guibas and Odlyzko for definitions) is approximately distributed as $\lfloor 4+\ln(n2^{-4}(1-2^{-4}))/\lambda+W/\lambda\rfloor$, where 4 is the length of the initial word in the

run, and $\lambda = \ln(2)$. Note here that the length of the run is measured in numbers of letters; in the previous case, length was measured in full words.

Similar results are possible for the B^* runs of Guibas and Odlyzko, in which case the chance of starting a given B^* run based on words of length m is $m2^{-(m+1)}$. Some delicacy is required in arguing $(m+1)$-dependence, because in this case runs may overlap.

The work of Erdös and Révész is extended to Markov chains by Samarova [13]. Our approach also carries over for Markov dependent sequences as well for pure runs of letters or words, save that $N(n)$ now depends on the stationary measure of the chain.

References

1. Anderson, C.W.: Extreme value theory for a class of discrete distributions with applications to some stochastic processes. J. Appl. Probal. **7**, 99–113 (1970)
2. Barndorff-Nielsen, O.: On the rate of growth of the partial maxima of a sequence of independent and identically distributed random variables. Math. Scand. **9**, 383–394 (1961)
3. Boyd, D.W.: Losing runs in Bernoulli trials. Unpublished manuscript (1972)
4. Carrier, G.F., Crook, M., Pearson, C.E.: Functions of a complex variable. Theory and technique. New York: McGraw Hill Book Co. 1966
5. Deheuvels, P.: On the Erdös-Rényi theorem for random sequences and its relationships with the theory of runs and spacing. Probab. Th. Rel. Fields **70**, 91–115 (1985)
6. Erdös, P., Révész, P.: On the length of the longest head run, pp. 219–228 in Topics in Information Theory. Colloquia Math. Soc. J. Bolyai 16 Keszthely (Hungary) (1975)
7. Ferguson, T.S.: On the distribution of max and mex. Manuscript (1984)
8. Guibas, L.J., Odlyzko, A.M.: Long repetitive patterns in random sequences. Z. Wahrscheinlichkeitstheorie verw. Gebiete **53**, 241–262 (1980)
9. Kendall, M., Stuart, A.: The Advanced Theory of Statistics. 4th Ed. Vol. 1 London and Wycombe: Charles Griffin and Co. Ltd. 1977
10. O'Brien, G.L.: Path properties of successive sample minima from stationary processes. Z. Wahrscheinlichkeitstheorie verw. Gebiete **38**, 313–327 (1977)
11. Robbins, H., Siegmund, D.: On the law of the iterated logarithm for maxima and minima. Proc. 6th Berkeley Sympos. Math. Statist. Probab. Univ. Calif. **3**, 51–70 (1972)
12. Rogosinski, W.: Fourier Series. New York: Chelsea Publishing Company 1959
13. Samarova, S.S.: On the asymptotic behaviour of the maximal sojourn time of an ergodic Markov chain in a fixed state. Russian Math. Surveys **35**(6), 103–104 (1980)
14. Solov'ev, A.O.: A combinatorial identity and its application to the problem concerning the first occurrence of a rare event. Theory Probab. Appl. **11**, 276–282 (1966)
15. Watson, G.S.: Extreme values in samples from m-dependent stationary stochastic processes. Ann. Math. Statist. 798–800 (1954)

Received April 15, 1984; in revised form September 2, 1985