View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Theoretical Computer Science 346 (2005) 358–387

Theoretical Computer Science

www.elsevier.com/locate/tcs

Monotone runs of uniformly distributed integer random variables: A probabilistic analysis

Guy Louchard*

Département d'Informatique, Université Libre de Bruxelles, CP 212, Boulevard du Triomphe, B-1050 Brusselles, Belgium

Abstract

Using a Markov chain approach and a polyomino-like description, we study some asymptotic properties of monotone increasing runs of uniformly distributed integer random variables. We analyze the limiting trajectories, which after suitable normalization, lead to a Brownian motion, the number of runs, which is asymptotically Gaussian, the run length distribution, the hitting time to a large length k run, which is asymptotically exponential, and the maximum run length which is related to the Gumbel extreme-value distribution function. A preliminary application to DNA analysis is also given.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Monotone increasing runs; Markov chains; Polyominoes

1. Introduction

We consider i.i.d. integer random variables with the distribution $I(s)$ given by $p(i) = 1/s, i = 1 \dots s$. Assume that we have a sequence $x[i], i = 1 \dots n$, constructed from independent $I(s)$ random variables. A run is a maximal (contiguous) increasing *monotone* subsequence. A maximum is a position i such that $x[i-1] \leqslant x[i] > x[i+1]$. For instance, with $s = 4$, the sequence 23241321224421321242 has 10 runs, with a maximal run length of 5. A first approach to the runs analysis was given in Crescenzi et al. [5]. We observe that a Markov chain approach can lead to asymptotic results on the runs. In this paper, continuing an approach we had started in [15,17–20], we consider the runs either as a stochastic

* Tel./fax: +322 6505613.
E-mail address: louchard@ulb.ac.be.

0304-3975/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2005.08.015

process or as a polyomino. This allows the derivation of several asymptotic distributions of random variables and processes such as: the limiting trajectories, which after suitable normalization, lead to a Brownian motion, the number of runs, which is asymptotically Gaussian, the run length distribution, the hitting time to a large length k run, which is asymptotically exponential, and the maximum run length, which is related to the Gumbel extreme-value distribution function. Our results are in some sense similar to what we have obtained in [18] where we analyzed geometrically distributed random variables, and [19], where we considered Carlitz compositions.

Let us already mention that, as $s \to \infty$, all properties related to the *ranked* $I(s)$ random variables have the same distributions as in the model of random permutations.

For more details on the analysis of algorithms on sequences, the interested reader should consult Szpankowski [23].

The paper is organized as follows: in Section 2, we present the associated Markov chain and the limiting trajectories. Section 3 is devoted to some simulations and the maximum run length distribution. Section 4 provides a preliminary application to DNA analysis. Section 5 concludes the paper. Appendix A–C provide some technical tools which are useful in Section 3.1.

This work emaned from the fruitful discussions we had with A. Del Lungo, during the 2003 summer. The paper was nearly completed when Alberto suddenly died on June 1, 2003. We present this work as a tribute to his memory.

2. Associated Markov chain and limiting trajectories

In this section, we first describe the runs process as a Markov chain. We then study the run length distribution. We finally analyze the limiting trajectories and the number of runs distribution.

2.1. Stochastic analysis

It is immediately checked that we can see the runs as a Markov chain, the states of which are given by the possible values i of the $I(s)$ random variables starting a run, together with the run length. The following notations will be used in the sequel:

$$
\begin{aligned}
L(i) &:= \text{length of an ascending run starting with value } i,\\
\varphi(i,l,j) &:= \Pr[L(i)=l, x[l+1]=j]\\
&= \sum_{i\leqslant i_2\leqslant\cdots\leqslant i_l>j} \left(\frac{1}{s}\right)^l,\\
\varphi(i,l) &:= \Pr[L(i)=l] = \sum_j \varphi(i,l,j)\\
&= \sum_{i\leqslant i_2\leqslant\cdots\leqslant i_l} \left(\frac{1}{s}\right)^{l-1} \frac{i_l-1}{s}, \quad \text{note that}\\
\sum_l \varphi(i,l) &= 1,
\end{aligned}
$$

$$\begin{aligned}\Pi\big[[i,l],[j,\lambda]\big] &:= \Pr[\text{next run starts at } j,\ \text{has length } \lambda,\\ &\qquad \text{given that the previous run was starting at } i\\ &\qquad \text{and had length } l]\\ &= \varphi(i,l,j)\varphi(j,\lambda)/\varphi(i,l).\end{aligned}$$

Two other transition matrices are given by

$$\begin{aligned}\Pi\big[[i,l],j\big] &:= \Pr[\text{next run starts at value } j,\\ &\qquad \text{given that the previous run was starting at } i\\ &\qquad \text{and had length } l]\\ &= \sum_{\lambda} \Pi\big[[i,l],[j,\lambda]\big]\\ &= \varphi(i,l,j)/\varphi(i,l),\end{aligned}$$

hence

$$\begin{aligned}\Pi\big[[i,l],[j,\lambda]\big] &= \Pi\big[[i,l],j\big]\varphi(j,\lambda),\\ \Pi[i,j] &:= \Pr[\text{next run starts at } j,\\ &\qquad \text{given that the previous run was starting at } i]\\ &= \sum_{l} \varphi(i,l,j).\end{aligned}$$

The stationary distribution of $\Pi[i,j]$ will be denoted by $\pi(j)$.

Note that $\varphi(i,l)$ can be computed directly. Indeed set, [1]

$$\begin{aligned}\varphi_1(i,l) &:= \sum_{w\geqslant l}\varphi(i,w) = \sum_{i\leqslant i_2\leqslant\cdots\leqslant i_l}\left(\frac{1}{s}\right)^{l-1}\\ &= [z^{l-1}]\left(\frac{1}{1-z/s}\right)^{s-i+1} = \frac{1}{s^{l-1}}\binom{s-i+l-1}{l-1}.\end{aligned}\tag{1}$$

We derive

$$\varphi(i,l) = \varphi_1(i,l) - \varphi_1(i,l+1) = \frac{1}{s^{l-1}}\binom{s-i+l-1}{l-1}\left(1-\frac{s-i+l}{ls}\right).\tag{2}$$

Also, for further use, we define

$$\begin{aligned}\varphi_2(i,l) &:= \sum_{w\geqslant l}\varphi_1(i,w),\\ \varphi_3(i,l) &:= \sum_{w\geqslant l}\varphi_2(i,w).\end{aligned}$$

These expressions are easily seen to converge. Also, by Stirling formula, or better, by singularity analysis, (see [9]), with the well-known formula

$$[z^n]\left(\frac{1}{1-z}\right)^{\alpha} \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right),\quad n\to\infty,$$

[1] We use the notation $[z^k]f(z)$ to denote the coefficient of z^k in the power expansion of $f(z)$.

we obtain

$$\varphi_1(i,k) \sim \frac{(k-1)^{s-i}}{(s-i)!s^{k-1}}\left[1+\mathcal{O}\left(\frac{1}{k}\right)\right]. \tag{3}$$

Let us denote by $F_j(z)$ the generating function of $\varphi_j(i,l)$. We know already that

$$F_1(i,z) = z\left(\frac{1}{1-z/s}\right)^{s-i+1}. \tag{4}$$

The generating function of $\varphi_2(i,l)$ is given by

$$F_2(i,z) = z\frac{F_1(i,z)-F_1(i,1)}{z-1}$$

and, by singularity analysis (the dominant singularity of F_2 is at $z=s$), we obtain

$$\varphi_2(i,k) \sim \frac{(k-1)^{s-i}}{(s-1)(s-i)!s^{k-2}}\left[1+\mathcal{O}\left(\frac{1}{k}\right)\right]. \tag{5}$$

Similarly, we derive

$$\varphi_3(i,k) \sim \frac{(k-1)^{s-i}}{(s-1)^2(s-i)!s^{k-3}}\left[1+\mathcal{O}\left(\frac{1}{k}\right)\right]. \tag{6}$$

2.2. Some generating functions and their applications

In this subsection, we first construct the generating functions of some probability transition matrices and stationary distributions. Next, we compute some stationary parameters distributions related to run lengths. Then we analyze the asymptotics of m successive runs.

2.2.1. One ascending run

We build a corresponding polyomino as follows: we choose a starting integer i, and construct a monotone sequence of l integers, $i = i_1 \leqslant \cdots \leqslant i_l$, followed by a last integer j with $j < i_l$. We mark each integer by z, the last monotone integer i_l is marked by η, and j is marked by θ. (For this kind of construction, see, for instance, [16].)

The trivariate generating function $\Psi(\eta,\theta,z|i)$ that comprises all such polyominoes, labelled as described, can be computed as follows: set

$$\begin{aligned} g_1 &:= z\eta^i, \quad \text{and, for } k \geqslant 1, \\ g_{k+1}(\eta,\theta,z) &= \sum_{v=1}^{s}[\eta^v]g_k(\eta,0,z)z\left[\sum_{r=v}^{s}\eta^r + \sum_{j=1}^{v-1}\theta^j\right] \\ &= g_k(1,0,z)z\theta\frac{1-\theta^s}{1-\theta} \\ &\quad + \sum_{v=1}^{s}[\eta^v]g_k(\eta,0,z)z\left[\eta^v\frac{\eta^{s-v+1}-1}{\eta-1} - \theta^v\frac{\theta^{s-v+1}-1}{\theta-1}\right]. \end{aligned}$$

This sequence describes the effect of "adding a new slice." The generating function g_k corresponds to a polyomino of k integers (each integer marked by z), where the last integer is marked either by η, if it is still monotone, or by θ, if it is smaller than the previous integer. Hence

$$\begin{aligned}\Psi(\eta,\theta,z|i) &:= \sum_{k=1}^{\infty} g_k(\eta,\theta,z)\\ &= g_1 + \Psi(1,0,z|i)z\theta\,\frac{1-\theta^s}{1-\theta} + z\,\frac{\eta^{s+1}}{\eta-1}\Psi(1,0,z|i)\\ &\quad -\frac{z}{\eta-1}\,\Psi(\eta,0,z|i) - \frac{\theta^{s+1}}{\theta-1}\,\Psi(1,0,z|i) + \frac{z}{\theta-1}\Psi(\theta,0,z|i).\end{aligned}$$

Set $A_1 := \Psi(1,0,z|i)$. (To ease the notation, we usually drop the variables list, except when the context is not sufficiently clear). Setting $\theta = 0$, $\eta = 0$, we immediately check that $\Psi(0,0,z|i) = 0$. Set now $\theta = 0$. We obtain

$$\Psi(\eta,0,z|i) = z\left[\eta^i + \frac{\eta^{s+1}}{\eta-1}\,A_1\right] - \frac{z}{\eta-1}\,\Psi(\eta,0,z|i),$$

which gives

$$\Psi(\eta,0,z|i) = z\,\frac{\eta-1}{\eta-1+z}\left[\eta^i + \frac{\eta^{s+1}}{\eta-1}\,A_1\right] \tag{7}$$

and

$$\begin{aligned}\Psi(\eta,\theta,z|i) &= z\left[\eta^i + \frac{\eta^{s+1}}{\eta-1}\,A_1 + \frac{1}{\theta-1}\,\Psi(\theta,0,z|i)\right]\\ &\quad + \frac{A_1\theta z}{1-\theta} - \frac{z^2}{\eta-1+z}\left[\eta^i + \frac{\eta^{s+1}}{\eta-1}A_1\right].\end{aligned} \tag{8}$$

We also derive from (7)

$$\Psi(\theta,0,z|i) = \frac{z(\theta-1)}{\theta-1+z}\left[\theta^i + \frac{\theta^{s+1}}{\theta-1}\,A_1\right]$$

and, finally

$$\begin{aligned}\Psi(\eta,\theta,z|i) &= z\left[\eta^i + \frac{\eta^{s+1}}{\eta-1}\,A_1\right] - \frac{z^2}{\eta-1+z}\left[\eta^i + \frac{\eta^{s+1}}{\eta-1}\,A_1\right]\\ &\quad + \frac{z^2}{\theta-1+z}\left[\theta^i + \frac{\theta^{s+1}}{\theta-1}\,A_1\right] + \frac{A_1\theta z}{1-\theta}\\ &= \frac{z(\eta-1)}{\eta-1+z}\left[\eta^i + \frac{\eta^{s+1}}{\eta-1}\,A_1\right]\\ &\quad + \frac{z^2}{\theta-1+z}\left[\theta^i + \frac{\theta^{s+1}}{\theta-1}\,A_1\right] + \frac{A_1\theta z}{1-\theta}.\end{aligned} \tag{9}$$

It remains to compute A_1. But the first part of (9) has a singularity in the denominator at $\eta = 1 - z$. This leads to

$$\left[\eta^{*i} + \frac{(1-z)^{s+1}}{1-z-1} A_1\right] = 0$$

or

$$A_1 = z(1-z)^{i-s-1} \quad \text{(which leads again to (1))},$$

and Ψ is completely determined.

Set now

$$\begin{aligned} \Gamma(\theta, z|i) &= \Psi(0, \theta, z|i) \\ &= \frac{A_1 \theta z}{1-\theta} + \frac{z^2}{\theta - 1 + z}\left[\theta^i + \frac{\theta^{s+1}}{\theta - 1} A_1\right] \end{aligned} \tag{10}$$

and, as $\theta \to 1$, this leads to

$$\Gamma(1, z|i) = z + (zs - 1)A_1.$$

The generating function Γ "contains" the quantities of interest, as described in the next proposition which gives all useful distributions in terms of Γ

Proposition 1.

$$\begin{aligned} \varphi(i, l, j) &= \frac{1}{s^l}\left[z^{l+1}\theta^j\right]\Gamma(\theta, z|i), \\ \varphi(i, l) &= \frac{1}{s^l}\left[z^{l+1}\right]\Gamma(1, z|i), \\ \Pi[i, j] &= \left[\theta^j\right] s\Gamma(\theta, 1/s|i). \end{aligned}$$

It is convenient to set

$$\begin{aligned} A_2(\kappa|i) &:= (1 - \kappa/s)^{i-s-1}, \\ A_3(\theta, \kappa|i) &:= A_2 \frac{\theta\kappa/s}{1-\theta} + \frac{\kappa/s}{\theta - 1 + \kappa/s}\left[\theta^i + \frac{\theta^{s+1}}{\theta - 1}\frac{\kappa}{s} A_2\right], \\ A_4(\kappa|i) &:= A_3(1, \kappa|i) = 1 + (\kappa - 1)A_2. \end{aligned}$$

This leads to

$$\begin{aligned} \varphi(i, l, j) &= [\kappa^l \theta^j] A_3(\theta, \kappa|i), \\ \varphi(i, l) &= [\kappa^l] A_4(\kappa|i) \quad \text{(which leads again to (2))}, \\ \Pi[i, j] &= [\theta^j] A_3(\theta, 1|i), \\ A_3(\theta, 1|i) &= (1 - 1/s)^{i-s-1} \frac{\theta}{s(1-\theta)} \\ &\quad + \frac{1}{s(\theta - 1 + 1/s)}\left[\theta^i + \frac{\theta^{s+1}}{s(\theta - 1)}(1 - 1/s)^{i-s-1}\right]. \end{aligned}$$

It is an easy check to derive that $\sum_j \Pi[i,j] = 1$, which follows from $\sum_j \Pi[i,j] = s\Gamma(1, 1/s|i)$.

To obtain $\pi(i)$, we start from the stationary equation

$$\pi(j) = \sum_i \pi(i)[\xi^j]A_3(\xi, 1|i).$$

We now compute the generating function of $\pi(i)$, $\Theta(\theta) := \sum_{j=1}^{s} \pi(j)\theta^j$, and $\Theta(1) = 1$

$$\Theta(\theta) = \sum_i \pi(i)A_3(\theta, 1|i) = (1-1/s)^{-s-1}\Theta((1-1/s))\frac{\theta}{s(1-\theta)} + \frac{1}{s(\theta-1+1/s)}\left[\Theta(\theta) + \frac{\theta^{s+1}}{s(\theta-1)}(1-1/s)^{-s-1}\Theta((1-1/s))\right]. \quad (11)$$

Solving (11) w.r.t. $\Theta(\theta)$ and letting $\theta \to 1$; we obtain

$$\Theta((1-1/s)) = 2(1-1/s)^s$$

and, finally,

$$\Theta(\theta) = 2\theta\,\frac{\theta^s - \theta s + s - 1}{s(1-\theta)^2(s-1)}, \quad (12)$$

from which we derive

$$\pi(i) = 2\,\frac{s-i}{s(s-1)}.$$

Of course, $\pi(s) = 0$ and $\sum_1^{s-1} \pi(i) = 1$.

2.2.2. Some stationary parameters distributions related to the run length

As in [18], we need to compute the following averages and second moments:

$$\bar{L}(i) = \sum_l \varphi(i,l)l,$$
$$\overline{L^2}(i) = \sum_l \varphi(i,l)l^2,$$
$$\bar{L} = \sum_i \pi(i)\bar{L}(i).$$

For instance (differentiation is w.r.t. κ), we obtain the following results:

Proposition 2.

$$\bar{L}(i) = A_4'\big|_{\kappa=1} = (1-1/s)^{i-s-1},$$
$$\overline{L^2}(i) = A_4''\big|_{\kappa=1} + \bar{L}(i) = \frac{(1-1/s)^{i-s}s(-2i+3s+1)}{(s-1)^2}, \quad (13)$$

and

$$\bar{L} = \sum \pi(i)\bar{L}(i) = 2/(1-1/s),$$
$$\overline{L^2} = \sum \pi(i)s^2(i) = \frac{2s(2(1-1/s)^{-s}-3)}{s-1}.$$

The expression for $\bar{L}$ is easy to check: this corresponds, by examining couples of random variables, to $s^2/\sum_{i=2}^{s}(i-1)$.

The stationary generating function of the run length is given by

$$H(\kappa) = \sum \pi(i)A_4 = 1 + (\kappa-1)(1-\kappa/s)^{-s-1}\Theta(1-\kappa/s), \tag{14}$$

from which we derive, for instance, the following first three values of the stationary run length distribution:

Proposition 3.

$$\Pr[L=1] = \frac{s-2}{3s},$$
$$\Pr[L=2] = \frac{5s^2-s-6}{12s^2},$$
$$\Pr[L=3] = \frac{(11s-12)(s+1)(s+2)}{60s^3}.$$

The stationary run length probability is geometrically decreasing. Note that, as $s \to \infty$, $\bar{L} \sim 2$, $\overline{L^2} \sim 4e-6$. This is clear from the following result: as $s \to \infty$, our model turns into the model of random permutations, with the stationary expressions $\bar{L} = 2$, $\overline{L^2} = 4e-6$. (see [20] for details).

For further use, let us define the truncated mean run length, starting from i:

$$\bar{L}(i,l) := \sum_{w<l} w\varphi(i,w).$$

By partial summation and (2), we obtain

$$\bar{L}(i,k) = \bar{L}(i) - G(i,k), \quad \text{with } G(i,k) := \sum_{w>k-1} w\varphi(i,w),$$
$$\begin{aligned} G(i,k) &= \sum_{l>k-2}(l+1)\varphi_1(i,l+1) - \sum_{l>k-1} l\varphi_1(i,l+1) \\ &= (k-1)\varphi_1(i,k) + \varphi_2(i,k) \\ &= k\varphi_1(i,k) + \varphi_2(i,k+1). \end{aligned} \tag{15}$$

A next useful quantity is given, with (12) and (4), by

$$H_1(k) := \sum \pi(i)\varphi_1(i,k) = [z^{k-1}]\left(\frac{1}{1-z/s}\right)^{s+1} \Theta(1-z/s) \tag{16}$$

$$= [z^{k+1}]2s\left[1+(z-1)\left(\frac{1}{1-z/s}\right)^{s}\right]\Big/(s-1)$$

$$= \frac{2}{(s-1)s^k}\left[s\binom{s+k-1}{k} - \binom{s+k}{k+1}\right]$$

$$= \frac{2k}{(s-1)s^k}\binom{s+k-1}{s-2}. \tag{17}$$

Asymptotically, singularity analysis leads to

$$H_1(k) \sim \frac{2}{(s-1)!s^k}k^{s-1}\left[1+\mathcal{O}\left(\frac{1}{k}\right)\right], \quad k\to\infty. \tag{18}$$

A last useful equivalent concerns

$$\sum_{l\geq k}\varphi(i,l,j)/\varphi_1(i,k), \quad k\to\infty,$$

i.e. the asymptotic distribution of a run starting value, J, after a long run. Setting

$$A_5(\theta,\kappa|i) := \frac{\kappa(A_3(\theta,\kappa|i)-A_3(\theta,1|i))}{\kappa-1},$$

we see that

$$\sum_{l\geq k}\varphi(i,l,j) = [\kappa^k\theta^j]A_5(\theta,\kappa|i).$$

But it is readily checked that neither $\kappa = s(1-\theta)$ nor $\kappa = 1$ are singularities of A_5, the only singularity is at $\kappa = s$. Hence we obtain the equivalent

$$A_5(\theta,\kappa|i) \sim \frac{s(\theta^s-\theta)}{(s-1)(\theta-1)}(1-\kappa/s)^{i-s-1}, \quad \kappa\to s.$$

Now,

$$\varphi_1(i,k) \sim [\kappa^{k-1}]\kappa(1-\kappa/s)^{i-s-1}, \quad k\to\infty$$

and the generating function of J is asymptotically given by

$$\frac{\theta^s-\theta}{(s-1)(\theta-1)}\left[1+\mathcal{O}\left(\frac{1}{k}\right)\right]. \tag{19}$$

But this intuitively evident: the long runs are followed by an integer asymptotically uniformly distributed on $[1\ldots s-1]$. We could continue the asymptotics: the next correction

term is given by a matrix (we do not need its explicit value here):

$$\varepsilon N(i,\, j) \tag{20}$$

with $\varepsilon := ((k-1)^{s-1})/s^{k-1}$ (the choice of ε is justified in Section 3.1). Note that $\mathbf{N1} = 0$.

2.2.3. m ascending runs

We will analyze the total length $X(m)$ (number of $I(s)$ random variables) of m runs. We have

$$X(m) = \sum_{k=1}^{m} L(i_k),$$

where i_k is the starting value of run k. This corresponds to the length of a word where each letter is a $I(s)$ random variable.

By the ergodic theorem, we know that

$$\begin{aligned}\mathbb{E}(X(m)) &\sim m\bar{L},\\ \mathrm{VAR}(X(m)) &\sim m\left[\overline{L^2} - \bar{L}^{\,2}\right] + 2mS = m\sigma^2, \text{ say}\end{aligned}$$

(see, for instance, [3, p. 172]), where

$$\begin{aligned}S &:= \sum_{v=1}^{\infty} \sum_{i,l,j,\lambda} \pi(i)\varphi(i,l)(l-\bar{L})\Pi^{v}[[i,l],[j,\lambda]](\lambda-\bar{L})\\ &= \sum_{v,i,l} \pi(i)\varphi(i,l)(l-\bar{L}) \sum_{w_1,w_2,\ldots,w_{v-1}} \Pi[[i,l],w_1]\Pi[w_1,w_2]\cdots\Pi[w_{v-2},w_{v-1}]\\ &\quad\times \sum_{j} \Pi[w_{v-1},j](\bar{L}(j)-\bar{L}).\end{aligned}$$

To compute S, we first analyze

$$f_0(\theta) := \sum_{k,i,l} \pi(i)\varphi(i,l,k)l\theta^k = \sum \pi(i)\; A_3'\big|_{\kappa=1}.$$

To simplify further expressions, we set $\theta = \eta(1-1/s)$. This leads to

$$\tilde{f}_0(\eta) = \frac{2\eta}{s((1-1/s)\eta-1)}\left[\left(\eta\frac{\eta^{s-1}-1}{\eta-1}+s\right)\Big/\,(s-1)-(1-1/s)^{-s}\right].$$

The first term of S is related to

$$\sum_{k}[\eta^k]\tilde{f}_0(\eta)\bar{L}(k)/(1-1/s)^k = (1-1/s)^{-s-1}\tilde{f}_0(1).$$

Now define

$$\begin{aligned}\tilde{f}_1(\eta) &:= \sum_{k,i,l,j} \pi(i)\varphi(i,l,k) l \Pi(k,j)\theta^j \\ &= \frac{(1-1/s)^{-s}\eta}{s(1-\eta(1-1/s))}\tilde{f}_0(1) \\ &\quad + \frac{1}{s(1-1/s)(\eta-1)}\left[\tilde{f}_0(\eta) + \frac{\eta^{s+1}}{s(\eta(1-1/s)-1)}\tilde{f}_0(1)\right],\end{aligned}$$

by (11). The second term of S is now related to $(1-1/s)^{-s-1}\tilde{f}_1(1)$.

We see now how to proceed: S is given by

$$S = \lim_{w\to 1}\left[(1-1/s)^{-s-1}\tilde{\phi}(w,1) - \bar{L}^2/(1-w)\right],$$

where

$$\tilde{\phi}(w,\eta) := \sum_0^\infty w^i \tilde{f}_i(\eta).$$

$\tilde{\phi}(w,\eta)$ can be computed as follows. We obtain

$$\begin{aligned}\tilde{\phi}(w,\eta) &= \tilde{f}_0(\eta) + \frac{w(1-1/s)^{-s}\eta}{s(1-\eta(1-1/s))}\tilde{\phi}(w,1) \\ &\quad + \frac{w}{s(1-1/s)(\eta-1)}\left[\tilde{\phi}(w,\eta) + \frac{\eta^{s+1}}{s(\eta(1-1/s)-1)}\tilde{\phi}(w,1)\right]\end{aligned}$$

or

$$\begin{aligned}&\tilde{\phi}(w,\eta)\frac{(1-1/s)(\eta-1)-w/s}{(1-1/s)(\eta-1)} \\ &\quad = \tilde{f}_0(\eta) + \frac{w\eta}{s((1-1/s)\eta-1)}\left[\frac{\eta^s}{s(1-1/s)(\eta-1)} - (1-1/s)^{-s}\right]\tilde{\phi}(w,1).\end{aligned}$$

To compute $\tilde{\phi}(w,1)$, we note that the denominator of the expression for $\tilde{\phi}(w,\eta)$ has a singularity at $\eta = 1 + w/(s(1-1/s))$, hence the numerator must be 0 for this value, which leads to

$$\tilde{\phi}(w,1) = 2\,\frac{-(1-1/s)^s + w(1-1/s)^s + ((s+w-1)/s)^s - w}{w[w-((s+w-1)/s)^s]}.$$

This gives

$$\begin{aligned}(1-1/s)^{-s-1}\tilde{\phi}(w,1) &= \bar{L}^2/(1-w) \\ &\quad - \frac{2s[3(1-1/s)^{-s}(s-1)-8s+4]}{3(s-1)^2} + \mathcal{O}(1-w),\end{aligned}$$

and the singularity (at $w = 1$) in S is removed. Finally,

$$\sigma^2 = \left[\overline{L^2} - \bar{L}^2\right] + 2S = \frac{2s(s+1)}{3(s-1)^2}.$$

Again, as $s \to \infty$, we obtain the limiting value $2/3$ which corresponds to a permutation.

2.3. Limiting trajectories

Let us denote by M the number of runs. Let us first fix M to m and let us consider the total length $X(m)$ (the number of $I(s)$ random variables, i.e. the length of the word) of m runs

$$X(m) = \sum_{k=1}^{m} L(i_k),$$

where $i_1, \ldots, i_m$ are the successive starting values of the runs. We can check that our Markov chain is φ-mixing (see, for instance, [3, p. 168] and the Appendix) and we apply the Functional Central Limit Theorem [3, p. 174, Theorem 20.1]. We obtain the following result, where $B(t)$ is the standard Brownian Motion and $\Rightarrow$ denotes the weak convergence of random functions in the space of all right-continuous functions that have right limits and are endowed with the Skorohod metric. This gives the limiting trajectories corresponding to a fixed number of runs.

Theorem 4.

$$\frac{X(\lfloor mt \rfloor) - \bar{L}mt}{\sigma\sqrt{m}} \Rightarrow B(t), \quad m \to \infty, \quad t \in [0, 1]. \tag{21}$$

Let us now condition on $X(m) = n$. M becomes a random variable. A realization of X for fixed n is given by Theorem 4, where we stop at a random time m such that $X(m) = n$. Proceeding as in [15] it is easy to check that this amounts to fix $m = n/\overline{L}$ in the denominator of (21), and we obtain the following result related to n random variables, with m runs.

Theorem 5. *Conditioned on* $X(m) = n$,

$$\frac{X(\lfloor mt \rfloor) - \bar{L}mt}{\sigma\sqrt{n/\bar{L}}} \Rightarrow B(t).$$

Moreover, from [22], the random variable $M :=$ number of runs satisfies the following Gaussian property:

Proposition 6.

$$M \sim \mathcal{N}(n\mu_1, n\sigma_1^2), \tag{22}$$

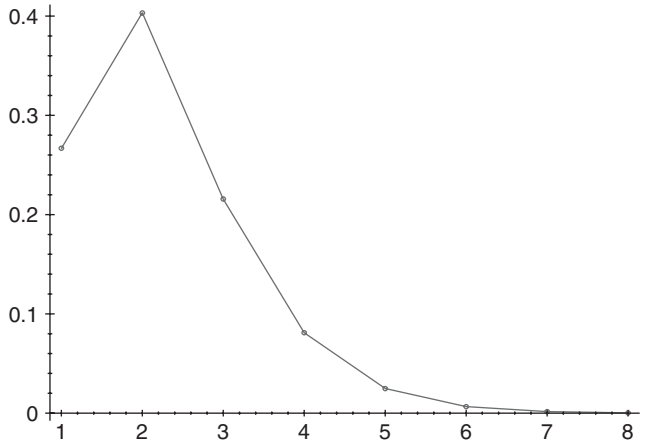

Fig. 1. Run length distribution (observed = circle, asymptotic = line).

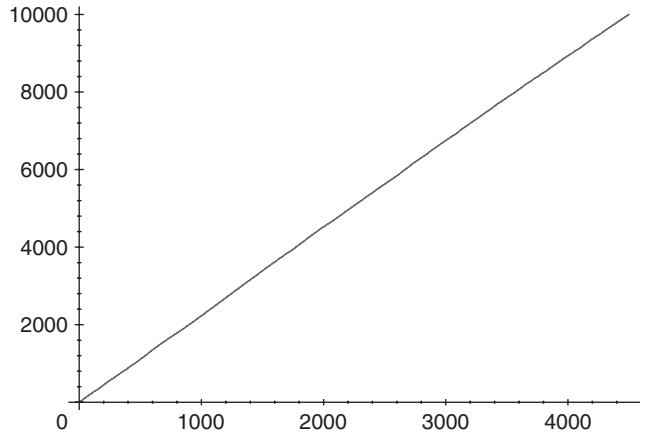

Fig. 2. Unnormalized $X(\cdot)$.

where

$$\mathcal{N} := a\ Gaussian\ (normal)\ random\ variable,$$
$$\mu_1 = 1/\bar{L} = \frac{s-1}{2s},$$
$$\sigma_1^2 = \sigma^2/\bar{L}^3.$$

This is actually an application of the generalized Renewal Theorem: the hitting time of the Brownian Motion $B(t)$ of (21) to the barrier $n(1-t)/(\sigma\sqrt{n/\bar{L}})$ is easily seen to lead asymptotically to a Gaussian random variable.

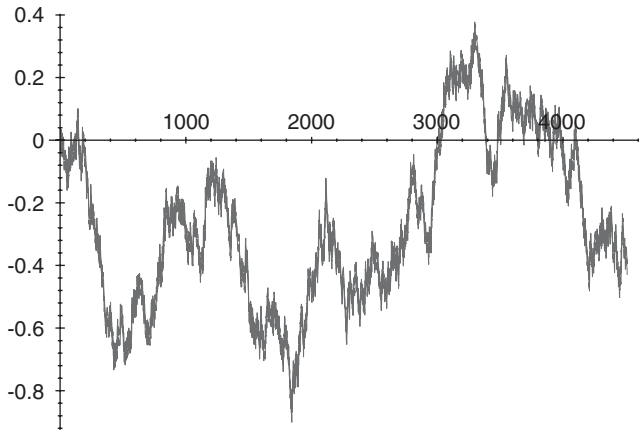

Fig. 3. Normalized $X(\cdot)$, with run length.

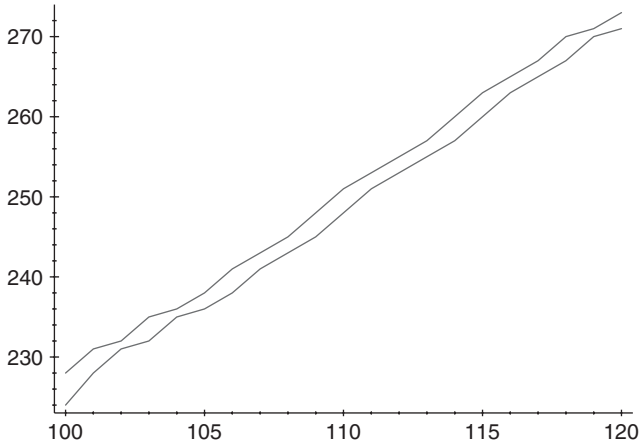

Fig. 4. Unnormalized $X(\cdot)$, with run length (zoom).

When $s \to \infty$, we obtain $\mu_1 \sim 1/2$, $\sigma_1^2 \sim 1/12$.

We have done a simulation of $U = 2000$ sequences of $n = 10\,000$ $I(10)$ random variables (this simulation will be extensively used in Section 3), leading to 9 002 231 runs. The observed and limiting run length distributions are given in Fig. 1. (The limiting distribution is given by Proposition 3). The fit is quite good. A typical trajectory for $X(i)$ for $n = 10\,000$, $s = 10$ has given $M = 4499$ runs. The unnormalized trajectory is given in Fig. 2, which shows a "filament silhouette." The normalized trajectory for $(X(i) - \bar{L}i)/\sigma\sqrt{M}$ and $(X(i-1) - \bar{L}i)/\sigma\sqrt{M}$, showing the run length is given in Fig. 3. Of course, both trajectories are asymptotically equivalent for $n \to \infty$. A zoom on $i = [100 \ldots 120]$ is given in Fig. 4. We have also checked the asymptotic Gaussian property of M as given by (22). This

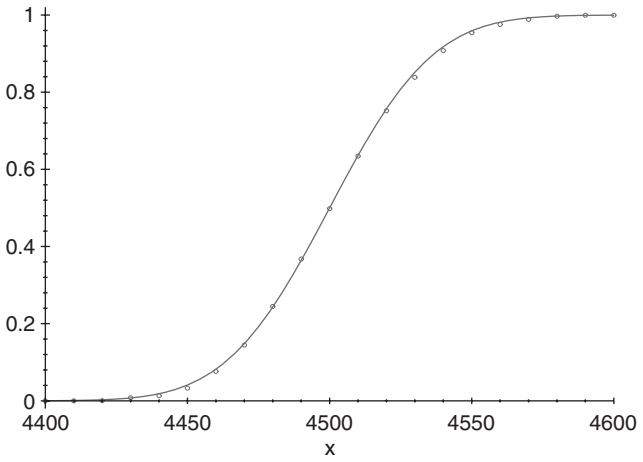

Fig. 5. Observed and limiting Gaussian M distribution function.

is shown in Fig. 5. The asymptotic mean and variance μ_1, σ_1^2 are given by 0.4500, $0.0825\ldots$ and the observed mean and variance by $0.4501\ldots$, $0.0840\ldots$.

3. Hitting time and maximum run length distribution

In this section, we study the hitting time to a large length k run and the maximum run length distribution. All vectors and matrices will be written in bold.

3.1. Large run in a sequence constructed from independent $I(s)$ random variables, for fixed s

Let us define $T_k :=$ the time (counted in terms of number of $I(s)$ random variables, i.e. the length of the word) necessary to obtain a run of large length k. We set $h_i := \mathbb{E}_i(T_k)$, (when possible, we drop the k for ease of notation). We derive

$$\begin{aligned} h_i &= [\bar{L}(i,k) + k\varphi_1(i,k)] + \sum_{l<k}\sum_u \varphi(i,l,u)h_u \\ &= [\bar{L}(i,k) + k\varphi_1(i,k)] + \sum_u \Pi(i,u)h_u - \sum_{l\geqslant k}\sum_u \varphi(i,l,u)h_u \\ &= [\bar{L}(i,k) + k\varphi_1(i,k)] + \sum_u \Pi(i,u)h_u \\ &\quad -\varphi_1(i,k)\sum_{l\geqslant k}\sum_u \frac{\varphi(i,l,u)}{\varphi_1(i,k)}h_u. \end{aligned} \tag{23}$$

By (3), we observe that $\varphi_1(i,k) = \mathcal{O}(\varepsilon)$, $\varepsilon := ((k-1)^{s-1})/s^{k-1}$, $k\to\infty$, uniformly on i, (for some reasons that will be clear later on, it is better to use the given form for ε), and

by standard properties of Markov chains (see [2,12]), we know that the hitting time to k is such that (we write D_i for $D_i(k)$):

$$h_i = \frac{D_i}{\varepsilon} + \psi(i) + \mathcal{O}(\varepsilon)$$

(Actually a Laurent series exists for ε sufficiently small),

$$\Pr_i[T_k \geqslant x] \sim \mathrm{e}^{-x/h_i}, \quad x \to \infty. \tag{24}$$

We will soon check that $\mathbf{D}$ is independent of i. First we note that, by (15) and (5),

$$\bar{L}(i,k) + k\varphi_1(i,k) = \bar{L}(i) + \mathcal{O}(\varepsilon).$$

Eq. (23), with (3), leads to

$$\begin{aligned} D_i + \varepsilon\psi(i) &= \varepsilon\bar{L}(i) + \left[\sum_u \Pi(i,u)[D_u + \varepsilon\psi(u)]\right] \\ &\quad -\varphi_1(i,k)\sum_{l\geqslant k}\sum_u \frac{\varphi(i,l,u)}{\varphi_1(i,k)}[D_u + \varepsilon\psi(u)] + \mathcal{O}(\varepsilon^2) \\ &= \mathbf{\Pi D}(i) + \mathcal{O}(\varepsilon). \end{aligned} \tag{25}$$

Identification of ε powers in (25) leads to $[\mathbf{I}-\mathbf{\Pi}]\mathbf{D} = 0$, which confirms that $\mathbf{D}$ is independent of i. Now we premultiply (23) by $\boldsymbol{\pi}$. This leads to

$$\boldsymbol{\pi}\mathbf{h} = \bar{L} + \boldsymbol{\pi}\mathbf{h} - \sum_{i=1}^{\infty} \pi(i)\varphi_1(i,k)\frac{D}{\varepsilon} + \mathcal{O}(\varepsilon)$$

or, with (17)

$$0 = \bar{L} - H_1(k)\frac{D}{\varepsilon} + \mathcal{O}(\varepsilon).$$

Set

$$G_3(k) := H_1(k)/\varepsilon = \frac{2k}{s(s-1)(k-1)^{s-1}}\binom{s+k-1}{s-2}. \tag{26}$$

This leads to

$$0 = \bar{L} - DG_3(k).$$

Therefore, we obtain

$$D = s^2(k-1)^{s-1} \Big/ \left[k\binom{s+k-1}{s-2}\right]. \tag{27}$$

We proceed now as in [17].

Let $\mathcal{M}(n) :=$ maximum run length based on n $I(s)$ random variables. We know that

$$\Pr[\mathcal{M}(n) < k] = \Pr[T_k \geqslant n+1].$$

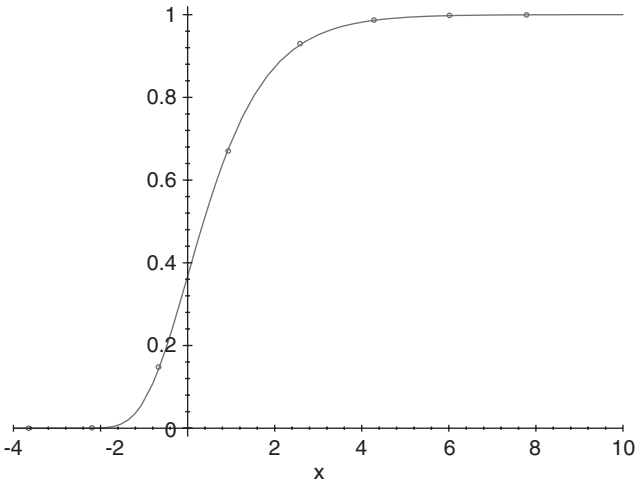

Fig. 6. Observed and limiting $\mathcal{M}$ distribution function.

We start our asymptotic analysis by first using (24)

$$\Pr[\mathcal{M}(n) < k] \sim \exp[-\exp[\log n + (s-1)\log(k-1) - (k-1)\log(s) - \log D]],$$

when $n \to \infty$.

Set now $k := j+1$ (which explains the choice of ε), $D(j) := s^2 j^{s-1} \Big/ \left[(j+1)\binom{s+j}{s-2}\right]$. We derive the following proposition, which gives the asymptotic distribution of the maximum run length.

Proposition 7. *Let*

$$\eta := j\ \log(s) - [\log n + (s-1)\log j - \log(D(j))].$$

Then, with integer j and $\eta = \mathcal{O}(1)$,

$$\Pr[\mathcal{M}(n) \leqslant j] \sim G_1(\eta), \quad n \to \infty,$$

where $G_1(\eta) := \exp[-\exp[-\eta]]$. *We also have*

$$\Pr[\mathcal{M}(n) = j] \sim f_1(\eta),$$

where $f_1(\eta) := G_1(\eta) - G_1(\eta - 1)$.

$G_1(.)$ is actually the Gumbel extreme-value distribution function. (See, for instance, Kotz and Nadarajah [14].) The observed and limiting distribution functions are given in Fig. 6 (observed = circle, asymptotic = line).

Let us make a few remarks:

- As $j \to \infty$, which is a consequence of $n \to \infty$, we have, by (18), $D(j) \to D_1 = s^2(s-2)!$.
  But $n = 10\,000$ is not large enough to entail this limit: in Fig. 6, we have kept the true $D(j)$ value.
- Again, as $n \to \infty$, $\log j \sim \log\log n - \log\log s$, but n must be large enough.
- We have a linear dependence on j in η. This is also found, for instance, in column-convex animals [17], Carlitz compositions [19], and runs of geometrically distributed random variables [18]. A quadratic dependence is found in ascending runs of sequences of geometrically distributed random variables [20], and in diagonally convex animals (see [17]), leading to a quite different behavior.

So we finally obtain the following theorem, which presents another form of the limiting distribution:

Theorem 8. *Let*

$$\eta_1 := j - \left[\log n + (s-1)(\log\log n - \log\log s) - \log(D_1)\right] / \log s.$$

Then, with integer j and $\eta_1 = \mathcal{O}(1)$,

$$\Pr[\mathcal{M}(n) \leqslant j] \sim G_2(\eta_1), \quad n \to \infty,$$

where $G_2(\eta_1) := \exp[-\exp[-\eta_1 \log s]]$. *Let*

$$\psi_1(n) := [\log n + (s-1)(\log\log n - \log\log s) - \log(D_1)]/\log s$$

and $\eta_1 = j - \lfloor\psi_1(n)\rfloor - \{\psi_1(n)\}$. *Asymptotically, the distribution is a periodic function of* $\psi_1(n)$ (*with period* 1), *which can be written as*

$$\log \Pr[\mathcal{M}(n) \leqslant \lfloor\psi_1(n)\rfloor + l]\mathrm{e}^{-\{\psi_1(n)\}\log s} \underset{n\to\infty}{\to} -\mathrm{e}^{-l\log s}.$$

We also have

$$\Pr[\mathcal{M}(n) = j] \sim f_2(\eta_1),$$

where $f_2(\eta_1) := G_2(\eta_1) - G_2(\eta_1 - 1)$.

3.2. The moments

If we had a rate of convergence property for Theorem 8, the asymptotic moments of $\mathcal{M}(n)$ would also be given by periodic functions of $\psi_1(n)$. They could be written as Harmonic sums which are usually analyzed with Mellin transforms: see Flajolet et al. [8]. The asymptotic non-periodic term in the moments of $\mathcal{M}(n)$ would be given by the following conjecture.

Conjecture 9. *The constant term* $\bar{\mathbb{E}}$ *in the Fourier expansion* (*in* $\psi_1(n)$) *of the moments of* $\mathcal{M}(n)$ *is asymptotically given by*

$$\bar{\mathbb{E}}[\mathcal{M}(n) - \psi_1(n)]^i \sim \int_{-\infty}^{+\infty} \eta^i [G_2(\eta_1) - G_2(\eta_1 - 1)]\, d\eta_1.$$

Indeed, the analysis of this type of Harmonic sums is detailed in Flajolet [7], where the author uses bounds for tail estimates. In our case, if our Markov chain were reversible, we would obtain, from Keilson [12], that

$$\left|\Pr_i[T_k \geqslant x] \sim \mathrm{e}^{-x/h_i}\right| \leqslant C\left[\frac{h_i^{(2)}}{2h_i^2} - 1\right]^{1/4}$$

for some constant C, uniformly on x and $h_i^{(2)} := \mathbb{E}_i(T_k^2)$. Now, by (A.1), (A.8) and (A.5),

$$\frac{h_i^{(2)}}{2h_i^2} - 1 = \mathcal{O}(\varepsilon).$$

For $k \log s$, respectively, given by $\log n - \log\log n$, $\log n$ and $2\log n$, the error bounds are $((\log n)^s/n)^{1/4}$, $((\log n)^{s-1}/n)^{1/4}$ and $(2\log n)^{(s-1)/4}/n^{1/2}$, which would be sufficient to establish our conclusions, following the lines of Flajolet [7]. It is well known that the extreme-value distribution function $\mathrm{e}^{-\mathrm{e}^{-x}}$ has mean γ and variance $\pi^2/6$. From this, we would for instance derive

$$\bar{\mathbb{E}}[\mathcal{M}(n)] \sim \psi_1(n) + \frac{1}{2} + \frac{\gamma}{\log s}.$$

The other periodic terms have very small amplitude (see [8]).

From Hitczenko and Louchard [11], we would also derive

$$\mathrm{VAR}[\mathcal{M}(n)] \sim \frac{\pi^2}{6(\log s)^2} + 1/12.$$

If we fix s to some given integer, then we can numerically derive a suitable rate of convergence. Indeed, from Aldous [1], we know that, if we define a time measure τ for the chain to approach stationarity:

$$\tau := \min\left\{n : \left[1/2\sum_{j=1}^{s-1}\left|\Pi^n[i,j] - \pi(j)\right|\right] \leqslant 1/e, \forall i\right\},$$

then

$$\sup_x\left|\sum_i \pi(i)\Pr_i[T_k \geqslant x] - \exp\left(-x\Big/\sum_i \pi(i)h_i\right)\right| \leqslant \Delta,$$

where

$$\Delta = C_1\tau\Big/\left[\sum_i \pi(i)h_i\left[1 + \log^+\left(\sum_i \pi(i)h_i/\tau\right)\right]\right]$$

and C_1 is a numerical constant.

As our chain is finite, τ is related to $1/|\lambda_2|$, where λ_2 is the second (in module) largest eigenvalue of Π. For instance,

for $s=4$, $\lambda_{2,1}=-7/81+4/81\sqrt{2}$, $\lambda_{2,2}=-7/81+4/81\sqrt{2}$, $|\lambda_2|=1/9$;
for $s=20$, $|\lambda_2|=0.1179696324....$

Here we have $h_i \sim D/\varepsilon$ and (C_i denote constants)

$$\Delta \sim C_2 \Big/ \left[\frac{D}{\varepsilon}\left\{1+\ln\left[C_3\,\frac{D}{\varepsilon}\right]\right\}\right]$$
$$\sim C_4\varepsilon/(-\ln(\varepsilon)) \sim C_5\varepsilon/(k\,\ln(s)) \sim C_6 k^{s-2}/s^k$$

and now, we have suitable error bounds and convergence of the moments. This is considered, in a much more general setting, in [21]

3.3. Maximum run length starting with value u

Let us now consider the maximum run length of type u (i.e. starting with value u). We derive

$$h_i(u) = \bar{L}(i) + \sum_j \Pi(i,j)h_j(u), \quad i \neq u,$$
$$h_u(u) = [\bar{L}(u,k) + k\varphi_1(u,k)] + \sum_j \Pi(u,j)h_j(u)$$
$$-\varphi_1(u,k)\sum_{l\geqslant k}\sum_j \frac{\varphi(u,l,j)}{\varphi_1(u,k)}\,h_j(u).$$

Proceeding as above, with now

$$\varepsilon := \frac{(k-1)^{s-u}}{(s-u-1)!s^{k-1}},$$

we obtain

$$D(u,j) = \frac{s^2(k-1)^{s-u}}{(s-u)!\binom{s-u+k-1}{k-1}},$$

and

$$\eta := j\log(s) - \left[\log n + (s-u)\log j - \log[(s-u-1)!] - \log(D(u,j))\right].$$

Note that, as $n\to\infty$, $D(u,j)\to s^2$. For $u=1$, the limiting distribution is given by Theorem 8. This means that the maximum run length is asymptotically only made of runs of type 1. For $u=2$, the simulation is given in Fig. 7.

3.4. Large run in a sequence constructed from independent $I(s)$ random variables, $s\to\infty$

Let us first consider (14). As $s\to\infty$, we obtain the stationary generating function

$$\lim_{s\to\infty} H(\kappa) = \left[2\kappa^2 e^{\kappa} + \kappa^2 - 4\kappa e^{\kappa} + 2\kappa + 2e^{\kappa} - 2\right]/\kappa^2,$$

which is exactly the corresponding formula for permutations (see [20]).

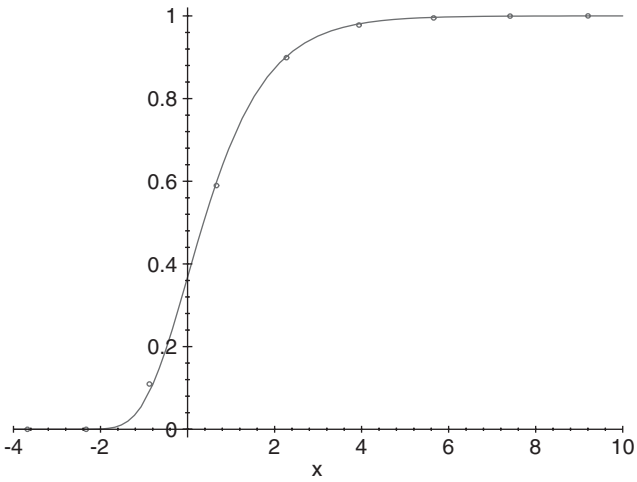

Fig. 7. Observed and limiting $\mathcal{M}$ distribution function, $u = 2$.

Also $\lim_{s\to\infty} \bar{L} = 2$ and, from (17), we easily derive, with Stirling,

$$\lim_{s\to\infty} H_1(k) = \frac{2k}{(k+1)!}.$$

Note that this can also be obtained from (16): as $s \to \infty$, this leads to a generating function

$$G_4(\kappa) = \frac{2(1+\kappa e^{\kappa} - e^{\kappa})}{\kappa^2},$$

and

$$[\kappa^{k-1}]G_4(\kappa) = \frac{2k}{(k+1)!}$$

which, again corresponds to the random permutation model: as in [20], we obtain

$$D/\varepsilon \sim \frac{(k+1)!}{k}.$$

4. Application: A preliminary analysis of a DNA sequence

As a simple example of DNA genome, we have first taken AL445469 (Human genomic DNA sequence from clone RP11-254N18), from which we have extracted a string S of $lt = 129\,480$ letters. We use the arbitrary transcription to integers: $a \to 1, c \to 2, g \to 3, t \to 4$. To obtain enough samples, we have divided S into 10 substrings of 12 948 integers each. Moreover, we have used all the $4! = 24$ permutations of letters order, leading to 240 strings, which give a reasonable estimate of the histograms we need.

First of all, on the complete string S, we have computed the observed distribution of integers, which is not uniform: this gives $p^*(1) = 0.34\ldots$, $p^*(2) = 0.19\ldots$, $p^*(3) = 0.18\ldots$, $p^*(4) = 0.29\ldots$. Moreover, the integers are not independent: an estimated

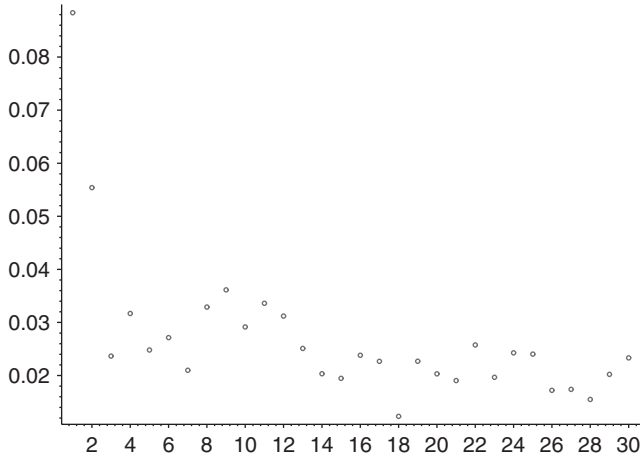

Fig. 8. Genome: observed correlation coefficient.

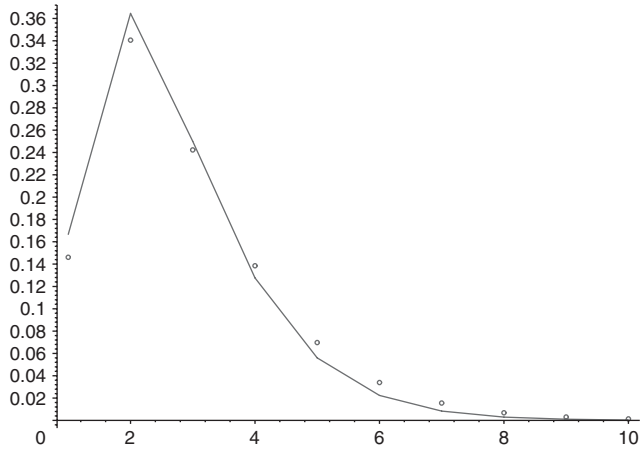

Fig. 9. Genome: limiting (line) and observed (circle) run length distribution.

correlation coefficient $\rho[k]$ on the string S is given in Fig. 8, with $\rho[1] = 0.088\ldots$, $\rho[2] = 0.055\ldots$. The other values are rather small. Next we have analyzed the run length: this leads to a total of 1 07 3964 runs. The fit with the limiting distribution is given in Fig. 9. The fit is reasonable.

Now we look at the maximal run length, based on the 240 substrings. The fit with the limiting distribution is given in Fig. 10, with a shift of 12 units!. The fit is very bad. Now we must try to analyze this discrepancy.

First we have checked whether it could come from the non-uniformity of the letters distribution. So we have generated lt independent, random integers with distribution p^*.

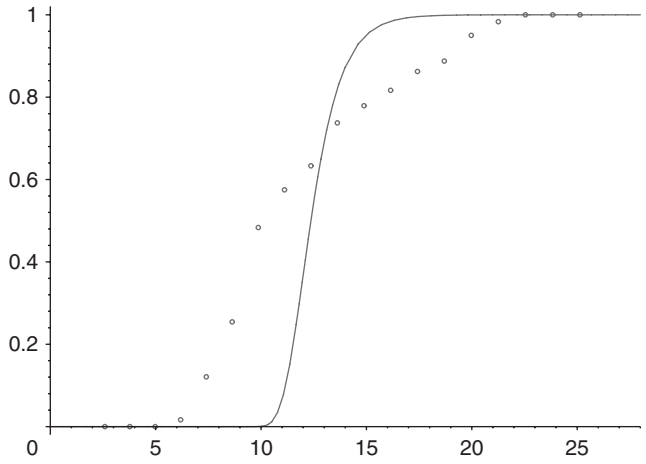

Fig. 10. Genome: limiting (line) and observed (circle) maximal run length distribution, shift = 12.

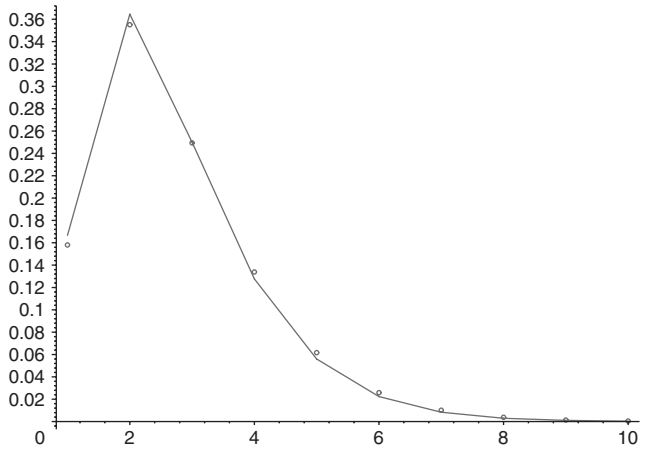

Fig. 11. i.i.d.: limiting (line) and observed (circle) run length distribution.

This leads to a total of 1 134 888 runs. The fit of the run length with the limiting distribution is given in Fig. 11. The fit is good: our asymptotic equivalents are robust.

Next we look at the maximal run length, based on the 240 substrings. The fit with the limiting distribution is given in Fig. 12, with a small shift of 0.3 units. The fit is rather good.

So the bad fit in the genome is not due to the non-uniformity. Then we have checked whether it could come from a first-order correlation. So we have computed from S a Markov chain $P[i, j]$ based on observed couples $S[k-1], S[k]$. We have then constructed a random string of lt random variables with this Markov chain. The observed first correlation coefficient is given by $\rho[1] = 0.082$. The fit of the run length with the limiting distribution

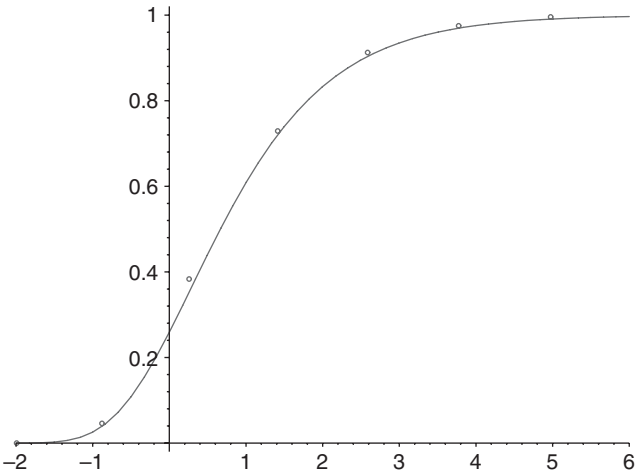

Fig. 12. i.i.d.: limiting (line) and observed (circle) maximal run length distribution, shift = 0.3.

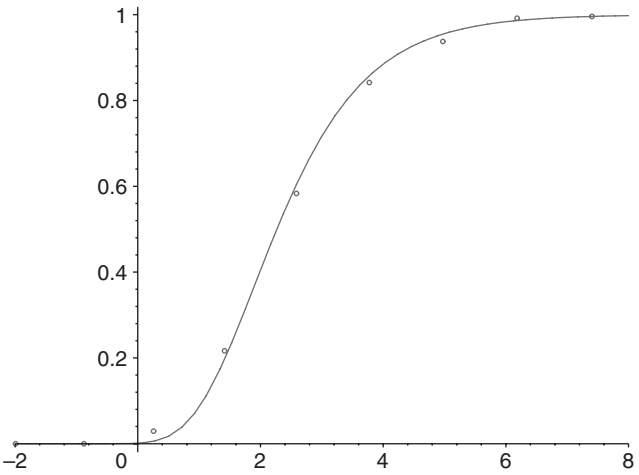

Fig. 13. Markov chain(order 1): limiting (line) and observed (circle) maximal run length distribution, shift = 1.9.

is again reasonable. The fit of the maximal run length with the limiting distribution is given in Fig. 13, with a shift of 1.9 units. The fit is rather good.

Finally, we have we have checked whether the discrepancy could come from a second-order correlation. So we have computed from S a Markov chain $P[i, j, l]$ based on observed triples $S[k-2]$, $S[k-1]$, $S[k]$ giving the probability that $S[k] = l$, given $S[k-2] = i$, $S[k-1] = j$. We have then constructed a random string of lt random variables with this Markov chain. The observed first two correlation coefficients are given by $\rho[1] = 0.064$, $\rho[2] = 0.028$. The fit of the run length with the limiting distribution is reasonable.

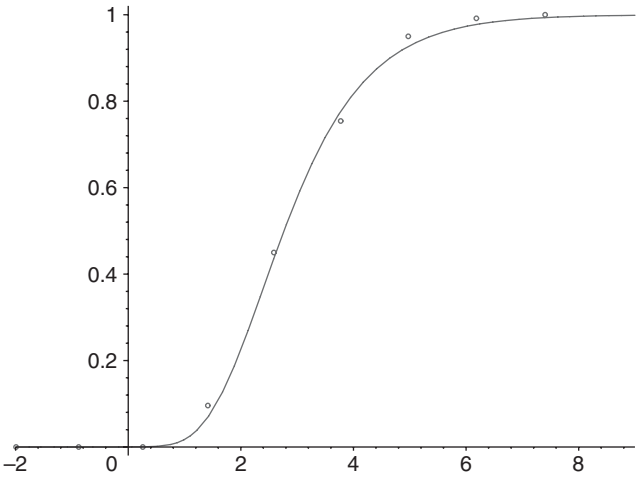

Fig. 14. Markov chain(order 2): limiting (line) and observed (circle) maximal run length distribution, shift $= 2.9$.

The fit of the maximal run length with the limiting distribution is given in Fig. 14, with a shift of 2.9 units. Again the fit is rather good.

So it seems that the discrepancy is due to some long range dependency, which is not observable from the correlation coefficients nor from the run length distribution, but discernable from the maximal run length. The biological explanation of this discrepancy will be investigated in some further work. Let us mention some related work: in [6], Denise et al. have analyzed the random generation of structured genomic sequences.

We have analyzed several other genomes: 6 470 333, 2 944 426, 2 627 293, 2 443 900, 15 620 559, 15 281 412. The results are very similar.

5. Conclusion

Using a Markov chain approach and a polyomino-like description, we have studied some asymptotic properties of monotone runs of uniformly distributed integer random variables. Some preliminary application to DNA analysis has revealed a long-range dependency. The biological explanation of this phenomena will be the object of some future work.

Acknowledgments

The author wishes to thank the Scienze Matematiche e Informatiche “Roberto Magari” for its warm hospitality. In particular, the cooperation with Alberto was quite agreeable and fruitful. He will be deeply regretted by his numerous friends.The careful reading and pertinent comments of the referee led to improvements in the presentation.

Appendix A. A detailed analysis of h_i and $h_i^{(2)}$

We must first be more precise in the analysis of h_i. With (3), (5), (6), we obtain, when $l = k + \mathcal{O}(1)$,

$$\varphi_1(i,l) = g_1(i,l)\varepsilon \quad \text{with } g_1(i,l) \sim \frac{(k-1)^{-i+1}}{s^{l-k}(s-i)!}\left[1+\mathcal{O}\left(\frac{1}{k}\right)\right],$$
$$\varphi_2(i,l) = g_2(i,l)\varepsilon \quad \text{with } g_2(i,l) \sim \frac{(k-1)^{-i+1}}{s^{l-k-1}(s-1)(s-i)!}\left[1+\mathcal{O}\left(\frac{1}{k}\right)\right],$$
$$\varphi_3(i,l) = g_3(i,l)\varepsilon \quad \text{with } g_3(i,l) \sim \frac{(k-1)^{-i+1}}{s^{l-k-2}(s-1)^2(s-i)!}\left[1+\mathcal{O}\left(\frac{1}{k}\right)\right].$$

Set

$$h_i = \frac{D(i)}{\varepsilon} + \psi(i) + \varepsilon\chi(i) + \mathcal{O}(\varepsilon^2). \tag{A.1}$$

We will now compute the dominant term of $\psi(i)$. We note that

$$\bar{L}(i,k) + k\varphi_1(i,k) = \bar{L}(i) - \varphi_2(i,k+1).$$

We obtain, from (23), (19), (20),

$$\begin{aligned} &D(i) + \varepsilon\psi(i) + \varepsilon^2\chi(i) \\ &\quad = \varepsilon\bar{L}(i) - \varepsilon^2 g_2(i,k+1) + \mathbf{\Pi D}(i) + \varepsilon\mathbf{\Pi}\boldsymbol{\psi}(i) + \varepsilon^2\mathbf{\Pi}\boldsymbol{\chi}(i) \\ &\qquad -\varphi_1(i,k)\left[D + \frac{1}{s-1}\sum_1^{s-1}\varepsilon\psi(u)\left[1+\mathcal{O}\left(\frac{1}{k}\right)\right] + \varepsilon\mathbf{ND}(i)\right] + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{A.2}$$

The ε term leads, for large k (apart from D, we neglect all $\mathcal{O}(1/k)$ error terms) to

$$\psi(i) = \bar{L}(i) + \mathbf{\Pi}\boldsymbol{\psi}(i) - Dg_1(i,k)$$

or

$$[\mathbf{I} - \mathbf{\Pi}]\boldsymbol{\psi} = \boldsymbol{\delta}, \tag{A.3}$$

where when $k \to \infty$,

$$\delta(1) = \bar{L}(1) - Dg_1(1,k),$$
$$\delta(i) = \bar{L}(i), \quad i > 1.$$

Now we premultiply (A.2) by $\boldsymbol{\pi}$. This leads, with G_3 defined by (26), to

$$\bar{L} = DG_3(k),$$

which conforms to (27). Next (A.3) leads to

$$\boldsymbol{\psi} = \mathbf{M}\boldsymbol{\delta} + \gamma_1, \quad \text{where} \quad \gamma_1 := \boldsymbol{\pi}\boldsymbol{\psi}$$

and $\mathbf{M} := \sum_{n\geqslant 0}(\mathbf{\Pi}^n - \mathbf{1}\times\boldsymbol{\pi})$. $\mathbf{M}$ is the Drazin inverse of $\mathbf{I}-\mathbf{\Pi}$. We refer to Campbell and Meyer [4] for a detailed definition and analysis of the Drazin inverse. We have $\mathbf{M} = \mathbf{Z} - \mathbf{1}\times\boldsymbol{\pi}$

where $\mathbf{Z} := [I - \mathbf{\Pi} + \mathbf{1} \times \boldsymbol{\pi}]^{-1} = \sum_{n \geqslant 0} [\mathbf{\Pi} - \mathbf{1} \times \boldsymbol{\pi}]^n$ is the potential used in Kemeny et al. [13]. First, by (B.1) (see the next appendix), we know that $\mathbf{M}\boldsymbol{\delta}$ is finite.

Next, to derive γ_1, we premultiply again (A.2) by $\boldsymbol{\pi}$. When $k \to \infty$, this leads to

$$D + \varepsilon \boldsymbol{\pi}\boldsymbol{\psi} + \varepsilon^2 \boldsymbol{\pi}\boldsymbol{\chi} = \varepsilon \bar{L} - \varepsilon^2 \pi(1) g_2(1, k+1) + D + \varepsilon \boldsymbol{\pi}\boldsymbol{\psi} + \varepsilon^2 \boldsymbol{\pi}\boldsymbol{\chi} - \varepsilon g_1(1,k)\pi(1) \left[D + \frac{1}{s-1} \sum_1^{s-1} \varepsilon \psi(u) + \mathcal{O}(\varepsilon^2) \right] + \mathcal{O}(\varepsilon^3).$$

The ε term leads of course again to D . The ε^2 term leads to

$$0 = -\pi(1) g_2(1, k+1) - g_1(1,k)\pi(1) \frac{1}{s-1} \sum_1^{s-1} \psi(u)$$

which allows the determination of γ_1.

We must now compute the first terms in the expansion of $h_i^{(2)} := \mathbb{E}_i(T_k^2)$. We derive

$$\begin{aligned} h_i^{(2)} &= k^2 \varphi_1(i,k) + \sum_{w<k} \varphi(i,w) w^2 \\ &\quad + 2 \sum_u \sum_w \varphi(i,w,u) w h_u + \sum_u \Pi(i,u) h_u^{(2)} \\ &\quad - 2 \sum_u \sum_{w \geq k} \varphi(i,w,u) w h_u - \sum_u \sum_{w \geq k} \varphi(i,w,u) h_u^{(2)}. \end{aligned} \tag{A.4}$$

We obtain, after some algebra, for large k,

$$k^2 \varphi_1(i,k) + \sum_{w<k} \varphi(i,w) w^2 = \overline{L^2}(i) + \varepsilon \left[-2k g_2(i,k+1) - 2 g_3(i,k+1) + g_2(i,k+1) \right] + \mathcal{O}(\varepsilon^2),$$

where $\overline{L^2}(i)$ is given by (13). We set

$$h_i^{(2)} = \frac{D^{(2)}(i)}{\varepsilon^2} + \frac{\psi^{(2)}(i)}{\varepsilon} + \chi^{(2)}(i) + \mathcal{O}(\varepsilon). \tag{A.5}$$

The next terms of (A.4) lead to (again, apart from D, we neglect all $\mathcal{O}(1/k)$ error terms)

$$\begin{aligned} &2\bar{L}(i) D/\varepsilon + 2 \sum_u \sum_w \varphi(i,w,u) w \psi(u) \\ &\quad + \mathbf{\Pi} \mathbf{D}^{(\mathbf{2})}(i)/\varepsilon^2 + \mathbf{\Pi}\boldsymbol{\psi}^{(\mathbf{2})}(i)/\varepsilon + \mathbf{\Pi}\boldsymbol{\chi}^{(\mathbf{2})}(i) \\ &\quad - 2DG(i,k)/\varepsilon - 2G(i,k) \frac{1}{s-1} \sum_1^{s-1} \psi(u) - 2\varepsilon G(i,k) \mathbf{N}^{(2)} \mathbf{1}(i) D/\varepsilon \\ &\quad - \varphi_1(i,k) \left[\frac{1}{s-1} \sum_1^{s-1} \left[D^{(2)}(u)/\varepsilon^2 + \psi^{(2)}(u)/\varepsilon \right] + \varepsilon \mathbf{N} \mathbf{D}^{(\mathbf{2})}(i)/\varepsilon^2 \right], \end{aligned}$$

where $G(i,k)$ is given by (15) and $\mathbf{N}^{(2)}$ is again a correction matrix related to $\sum_{w\geq k}\ \varphi(i,w,u)w$. Eq. (A.4) leads, for large k, to

$$\begin{aligned}
&D^{(2)}(i)+\varepsilon\psi^{(2)}(i)+\varepsilon^2\chi^{(2)}(i)\\
&\quad=\varepsilon^2\overline{L^2}(i)+2\varepsilon D\bar{L}(i)+2\varepsilon^2\sum_u\sum_w\varphi(i,w,u)w\psi(u)\\
&\qquad+\boldsymbol{\Pi}\mathbf{D}^{(\mathbf{2})}(i)+\varepsilon\boldsymbol{\Pi}\boldsymbol{\psi}^{(\mathbf{2})}(i)+\varepsilon^2\boldsymbol{\Pi}\boldsymbol{\chi}^{(\mathbf{2})}(i)-2\varepsilon D[\varphi_1(i,k)+\varphi_2(i,k+1)]\\
&\qquad-\varphi_1(i,k)\frac{1}{s-1}\sum_1^{s-1}\Big[D^{(2)}(u)+\varepsilon\psi^{(2)}(u)\Big]+\mathcal{O}(\varepsilon^3).
\end{aligned}\tag{A.6}$$

The constant term leads to $[\mathbf{I}-\boldsymbol{\Pi}]\mathbf{D}^{(\mathbf{2})}=0$, which confirms that $\mathbf{D}^{(\mathbf{2})}$ is independent of i. The ε term leads, for large k, to

$$\psi^{(2)}(i)=2\bar{L}(i)D+\boldsymbol{\Pi}\boldsymbol{\psi}^{(\mathbf{2})}(i)-D^{(2)}g_1(i,k)$$

or

$$[\mathbf{I}-\boldsymbol{\Pi}]\boldsymbol{\psi}^{(\mathbf{2})}=\boldsymbol{\delta}^{(\mathbf{2})},\tag{A.7}$$

where when $k\to\infty$, (again, apart from $D^{(2)}$, we neglect all $\mathcal{O}(1/k)$ error terms)

$$\begin{aligned}
&\delta^{(2)}(1)=2\bar{L}(1)D-D^{(2)}g_1(1,k),\\
&\delta^{(2)}(i)=2\bar{L}(i)D,\quad i>1.
\end{aligned}$$

Now we premultiply (A.6) by $\boldsymbol{\pi}$. This leads to

$$0=2\bar{L}D-D^{(2)}G_3(k).$$

Hence

$$D^{(2)}\equiv 2D^2\tag{A.8}$$

as it should. Next (A.7) leads to

$$\boldsymbol{\psi}^{(\mathbf{2})}=\mathbf{M}\boldsymbol{\delta}^{(\mathbf{2})}+\gamma_2,\quad\text{where}\quad\gamma_2:=\boldsymbol{\pi}\boldsymbol{\psi}^{(\mathbf{2})}.$$

First, by (B.1), we know that $\mathbf{M}\boldsymbol{\delta}^{(\mathbf{2})}$ is finite.

Next, to derive γ_2, we premultiply again (A.6) by $\boldsymbol{\pi}$. This leads to

$$\begin{aligned}
&D^{(2)}+\varepsilon\boldsymbol{\pi}\boldsymbol{\psi}^{(\mathbf{2})}+\varepsilon^2\boldsymbol{\pi}\boldsymbol{\chi}^{(\mathbf{2})}\\
&\quad=\varepsilon^2\overline{L^2}+2\varepsilon D\bar{L}+2\varepsilon^2\sum_i\pi(i)\sum_u\sum_w\varphi(i,w,u)w\psi(u)\\
&\qquad+D^{(2)}+\varepsilon\boldsymbol{\pi}\boldsymbol{\psi}^{(\mathbf{2})}+\varepsilon^2\boldsymbol{\pi}\boldsymbol{\chi}^{(\mathbf{2})}-2\varepsilon^2D\left[kG_3(k)+\sum_i\pi(i)g_2(i,k+1)\right]\\
&\qquad-\varepsilon G_3(k)D^{(2)}-\varepsilon^2G_3(k)\frac{1}{s-1}\sum_1^{s-1}\psi^{(2)}(u)+\mathcal{O}(\varepsilon^3).
\end{aligned}$$

The ε term leads of course again to $D^{(2)} = 2D^2$. The ε^2 term leads, when $k \to \infty$, to

$$0 = s^2 + 2\sum_i \pi(i) \sum_u \sum_w \varphi(i, w, u) w \psi(u) - 2D\pi(1)\,[k g_1(1, k) + g_2(1, k+1)]$$
$$-\pi(1) g_1(1, k) \frac{1}{s-1} \sum_1^{s-1} \psi^{(2)}(u),$$

which allows the determination of γ_2.

Appendix B. An explicit expression for M

We proceed as in Section 2.2.3. We must compute $\mathbf{M} := \sum_{n \geqslant 0} (\mathbf{\Pi}^n - \mathbf{1} \times \boldsymbol{\pi})$. We see that

$\mathbf{S} := \sum_{n \geqslant 1} (\mathbf{\Pi}^n - \mathbf{1} \times \boldsymbol{\pi})$ can be written as

$$S(i, j) = \lim_{w \to 1} \left[\sum_{n \geqslant 1} w^n \Pi^n(i, j) - \frac{w}{1-w} \pi(j) \right].$$

But this amounts to compute

$$S(i, j) = \lim_{w \to 1} \left[\alpha^i [\eta^j] (\tilde{\phi}(w, \eta) - \eta^i)/w + \frac{1}{1-w} \pi(j) \right],$$

where $\tilde{\phi}(w, \eta)$ (see Section 2.2.3) is now computed with the starting value $\tilde{f}_0(\eta) = \eta^i$. After all computations, and resubstituting $\eta = \theta/\alpha$, this gives a first term $\Theta(\theta)/(1-w)$ (see (12)), which cancels the singularity at $w = 1$, and finally, we obtain

$$S(i, j) = [\theta^j] R(\theta),$$

where

$$R(\theta) = -3\theta^i / (s(1-\theta) - \{2\theta^{s+1}[(-s-4+3i) + (1+s-3i)\theta]$$
$$+2\theta[[4+3i(s-1) - s(s+3)] + [-1+3i+s(2s-6i+4)]\theta$$
$$+s(3i-1-s)\theta^2]\}/[(s-1)s^2(1-\theta)^3].$$

And finally,[2]

$$M(i, j) = S(i, j) + [\![i = j]\!] - \pi(j). \tag{B.1}$$

Appendix C. The φ-mixing property

Let $\ldots, x_{-1}, x_0, x_1, \ldots$ be a strictly stationary sequence of random variables. For $a \leqslant b$, define $\mathcal{M}_a^b$ as the σ-field generated by the random variables $x_a, \ldots, x_b$. Consider a non-negative function φ of positive integers. We shall say that the sequence x_n is φ-mixing if,

[2] Here we use the indicator function notation proposed by Knuth et al. [10].

for each k, $-\infty < k < \infty$, and for each n, $n \geqslant 1$, $E_1 \in \mathcal{M}_{-\infty}^{k}$ and $E_2 \in \mathcal{M}_{k+n}^{\infty}$ together imply

$$|\Pr(E_1 \bigcap E_2) - \Pr(E_1 - \Pr(E_2)| \leqslant \varphi(n) \Pr(E_1).$$

If the Markov chain is finite, irreducible and aperiodic, then x_n is φ-mixing, with $\varphi(n) = a\rho^n$, $\rho < 1$.

References

[1] D. Aldous, Markov chains with almost exponential hitting times, Stochast. Process. Appl. 13 (1982) 305–310.
[2] D. Aldous, Probability Approximations via the Poisson Clumping Heuristic, Springer, Berlin, 1989.
[3] P. Billingsley, Convergence of Probability Measures, Wiley, New York, 1968.
[4] S.L. Campbell, C.D. Meyer, Generalized Inverse of Linear Transformations, Pitman, London, 1979.
[5] P. Crescenzi, A. Del Lungo, R. Grossi, E. Lodi, L. Pagli, G. Rossi, Text sparsification via local maxima, in: 20th Conf. on the Foundation of Software Technology and Theoretical Computer Science, 2000.
[6] A. Denise, Y. Ponty, M. Termier, Random generation of structured genomic sequences, in: Proc. RECOMB2003, 2003.
[7] P. Flajolet, Approximate counting: a detailed analysis, BIT 25 (1985) 113–134.
[8] P. Flajolet, X. Gourdon, P. Dumas, Mellin transforms and asymptotics: Harmonic sums, Theoret. Comput. Sci. 144 (1995) 3–58.
[9] P. Flajolet, B. Sedgewick, Analytic combinatorics, Singularity Analysis of Generating Functions, 2004, to appear (Chapter 6).
[10] R.L. Graham, D.E. Knuth, O. Patashnik, Concrete Mathematics, second ed., Addison-Wesley, Reading, MA, 1994.
[11] P. Hitczenko, G. Louchard, Distinctness of compositions of an integer: a probabilistic analysis, Random Structures Algorithms 19 (3,4) (2001) 407–437.
[12] J. Keilson, Markov Chain Models-Rarity and Exponentiality, Springer, Berlin, 1979.
[13] J.G. Kemeny, J.L. Snell, A.W. Knapp, Denumerable Markov Chains, Van Nostrand, Princeton, NJ, 1966.
[14] S. Kotz, S. Nadarajah, Extreme Value Distributions, Imperial College Press, 2000.
[15] G. Louchard, Probabilistic analysis of some directed animals, Theoret. Comput. Sci. 159 (1) (1996) 65–79.
[16] G. Louchard, Probabilistic analysis of column-convex and directed diagonally convex animals, Random Structures Algorithms 11 (1997) 151–178.
[17] G. Louchard, Probabilistic analysis of column-convex and directed diagonally-convex animals. II: trajectories and shapes, Random Structures Algorithms 15 (1999) 1–23.
[18] G. Louchard, Runs of geometrically distributed random variables: a probabilistic analysis, J. Comput. Appl. Math. 142 (1) (2002) 137–153.
[19] G. Louchard, H. Prodinger, Probabilistic analysis of Carlitz compositions, Discrete Math. Theoret. Comput. Sci. 5 (1) (2002) 71–96.
[20] G. Louchard, H. Prodinger, Ascending runs of geometrically distributed random variables: a probabilistic analysis, Theoret. Comput. Sci. 304 (2003) 59–86.
[21] G. Louchard, H. Prodinger, The moments problem of extreme-value related distribution functions, 2004, http://www.ulb.ac.be/di/mcs/louchard/mom7.ps.
[22] G. Louchard, W. Szpankowski, Average profile and limiting distribution for a phrase size in the Lempel–Ziv parsing algorithm, IEEE Trans. Inform. Theory 41 (1995) 478–488.
[23] W. Szpankowski, Average Case Analysis of Algorithms on Sequences, Wiley, New York, 2001.