Runs of geometrically distributed random variables: a probabilistic analysis

Guy Louchard *

March 13, 2001

Abstract

Using a Markov chain approach and a polyomino-like description, we study some asymptotic properties of runs of geometrically distributed random variables. We analyze the limiting trajectories, the number of runs and the run length distribution, the hitting time to a length k run and the maximum run length.

1 Introduction

During the LATIN 2000 Symposium in Montevideo, our attention was attracted to a talk given by A.Knopfmacher on the run statistics for geometrically distributed random variables (R.V.). The corresponding extended abstract by P.J.Grabner, A.Knopfmacher, and H.Prodinger [8] presented an interesting set of results on the asymptotic properties of runs. Let a GEOM(p) R.V. , with distribution $p(i) := pq^{i-1}, q := 1 - p$. The combinatorics of n i.i.d GEOM(p) R.V. has attracted recent interest : let us mention the skip list (P.Kirschenhofer, H.Prodinger, [12]) and the probabilistic counting (P.Kirschenhofer, H.Prodinger, and W.Szpankowski, [13]). Ascending runs statistics play a significant role in the behaviour of sorting algorithms (see D.E.Knuth [14]). We discovered that a Markov chain approach could lead to further asymptotic results on the runs of equal consecutive values.

In this paper, continuing an approach we had started in [15], [16], [17], we consider the runs either as a stochastic process or as a polyomino. This allows the derivation of several asymptotic distributions of R.V. and processes such as: limiting trajectories, number of runs and run length distribution, hitting time to a length k run and maximum run length.

The paper is organized as follows: in Sec.2, we present the associated Markov chain and the limiting trajectories. Sec.3 is devoted to some simulations and the maximum run length distribution. Sec.4 concludes the paper. An Appendix provides some technical tools which are necessary in Sec.3.2.

2 Associated Markov Chain and limiting trajectories

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

*Université Libre de Bruxelles, Département d'Informatique, CP 212, Boulevard du Triomphe, B-1050 Bruxelles, Belgium, email:louchard@ulb.ac.be

2.1 Stochastic analysis

Our probabilistic model is the following: we consider a sequence of i.i.d. GEOM(p) R.V. A run is defined as a sequence of equal consecutive values. 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 GEOM(p) R.V. The transition matrix is given by

$$\Pi(i,j) = p(j)/(1-p(i)), \quad j \neq i, \tag{1}$$

and the process stays, at each state i, for a length $L(i)$, with another GEOM($p(i)$) distribution:

$$\varphi(i,l) := \Pr[L(i) = l] = p(i)^{l-1}(1-p(i)). \tag{2}$$

The chain $\mathbf{\Pi}$ is irreducible, strongly ergodic, and reversible.
Several useful parameters are computed as follows.
The stationary distribution $\boldsymbol{\pi}$ of (1) is readily given by

$$\begin{aligned} \pi(i) &= \alpha p(i)(1-p(i)), \quad \text{with} \\ \alpha &= 1/\sum_{i=1}^{\infty} p(i)(1-p(i)) = \frac{1+q}{2q}. \end{aligned}$$

The first moments of $L(i)$ are given by

$$\begin{aligned} \bar{L}(i) &= \mathbb{E}(L(i)) = \sum_{l=1}^{\infty} l\varphi(i,l) = 1/(1-p(i)), \\ s^2(i) &= \sum_{l=1}^{\infty} l^2\varphi(i,l) = \frac{1+p(i)}{(1-p(i))^2}. \end{aligned} \tag{3}$$

The probability generating function (PGF) $F(i,z)$ of the run length is given by

$$F(i,z) = \sum_{l=1}^{\infty} z^l \varphi(i,l) = \frac{(1-p(i))z}{1-p(i)z}.$$

For further use, let us define

$$\Lambda(i,l) := \sum_{w \leq l} \varphi(i,w) = 1 - p(i)^l \tag{4}$$

and

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

Let us denote by M the number of runs . Let us first fix M to m and let us consider the total length $X(m)$ (number of GEOM R.V.) of m runs :

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

where $i_1 \ldots i_m$ are the successive GEOM values. By the ergodic theorem, we know that

$$\mathbb{E}(X(m)) \sim m\bar{L},$$

where

$$\bar{L} = \sum_{i=1}^{\infty} \pi(i)\bar{L}(i) = \alpha$$

and

$$\begin{aligned}
\mathbb{E}(X(m)^2) &\sim ms^2 + 2m\sum_{k=1}^{\infty} C_k, \\
s^2 &= \sum_{i=1}^{\infty} \pi(i)s^2(i) = \alpha \sum_{i=1}^{\infty} p(i)(1+p(i))/(1-p(i)), \qquad (5)\\
C_k &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \pi(i)\bar{L}(i)\Pi^k(i,j)\bar{L}(j).
\end{aligned}$$

2.2 Limiting trajectories

The computation of VAR$(X(m))$ can be done as follows. First we note that $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\pi(i)\bar{L}(i)\Pi(i,j)\bar{L}(j)$ can be written as

$$\alpha \sum_{i=1}^{\infty} \sum_{\substack{j=1 \\ j\neq i}}^{\infty} \frac{p(i)}{1-p(i)}\frac{p(j)}{1-p(j)}.$$

More generally, $\sum_i \sum_j \pi(i)\bar{L}(i)\Pi^{k-1}(i,j)\bar{L}(j)$ is given by

$$\alpha \sum_{i_1=1}^{\infty} \sum_{\substack{i_2=1 \\ i_2\neq i_1}}^{\infty} \cdots \sum_{\substack{i_k=1 \\ i_k\neq i_{k-1}}}^{\infty} \frac{p(i_1)}{1-p(i_1)}\frac{p(i_2)}{1-p(i_2)}\cdots\frac{p(i_k)}{1-p(i_k)}.$$

Also, $p(i)/(1-p(i))$ is given by

$$\frac{pq^i}{q}[1+\frac{pq^i}{q}+(\frac{pq^i}{q})^2+\ldots].$$

So, generalizing [16], [17], we build a polyomino as follows: we choose a starting integer i. We construct $L(i)$ copies of this integer. We call this a run. Then, we choose another integer $j \neq i$ and construct the corresponding run, etc. A polyomino of m runs has a total number of integers $X(m) = \sum_{k=1}^{m} L(i_k)$. Let us call ξ the total size of the polyomino, i.e. the sum of all integer values:

$$\xi := \sum_{k=1}^{m} L(i_k)i_k.$$

We mark ξ by x, X by z, m by w and the value of the last integer i_m by θ. The quadrivariate generating function(GF) is now given by (we omit the details, the technique is described in [16] and [17]),

$$\phi(w,\theta,x,z|i) = A_2(w,\theta,x,z|i) + A_1(w,\theta,x,z)H_2(w,x,z|i)$$

where

$$\begin{aligned}
H_2(w,x,z|i) &= A_2(w,1,x,z|i)/g(w,x,z), \\
g(w,x,z) &= 1 - A_1(w,1,x,z), \\
A_2(w,\theta,x,z|i) &= \frac{\theta^i w x^i z}{1-x^i z}\Big/\left[1+\frac{wx^iz}{1-x^iz}\right], \\
A_1(w,\theta,x,z) &= \sum_{i=1}^{\infty} A_2(w,\theta,x,z|i).
\end{aligned}$$

Now , we derive from (5)

$$\mathrm{VAR}(X(m)) \sim m[s^2 - \bar{L}^2] + 2mS$$

where

$$S := \sum_{k=1}^{\infty}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \pi(i)(\bar{L}(i) - \bar{L})\Pi^k(i,j)(\bar{L}(j) - \bar{L}).$$

It is now easy to check that S can be written as

$$\begin{aligned} S &= \lim_{w\to 1}\left[\alpha \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}[\theta^j]\left\{\frac{A_2(w,\theta,x^*,z^*|i) - w\theta^i x^{*i}z^*/(1-x^{*i}z^*)}{w}\right.\right. \\ &+ \left.\left.\frac{A_1(w,\theta,x^*,z^*)A_2(w,1,x^*,z^*|i)}{wg(w,x^*,z^*)}\right\} - \frac{\bar{L}^2 w}{1-w}\right] \end{aligned}$$

with $x^* = q, z^* = p/q$, which leads to $x^{*i}z^* \equiv p(i)$. Set $w = 1-\varepsilon$. We note that $g(1,x^*,z^*) = 0$. So we obtain

$$S = \lim_{\epsilon\to 0}\left[-\frac{\bar{L}^2}{\epsilon} + \bar{L}^2 + \alpha\bar{\varphi}_2 + \alpha\left[-\frac{\bar{\varphi}_1(1)}{g_w\epsilon} + \frac{\bar{\varphi}_{1,w}(1)}{g_w} - \frac{\bar{\varphi}_1(1)g_{ww}}{2g_w^2}\right]\right] \tag{6}$$

where

$$\begin{aligned} \bar{\varphi}_2 &= \sum_{i=1}^{\infty}\left[\frac{p(i)}{1-p(i)}\Big/\left[1+\frac{p(i)}{1-p(i)}\right] - \frac{p(i)}{1-p(i)}\right] = \sum_{i=1}^{\infty} -p(i)^2/(1-p(i)), \\ \varphi_1(w) &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} w^2\frac{p(i)}{1-p(i)}\frac{p(j)}{1-p(j)}\Big/\left\{w\left[1+\frac{wp(i)}{1-p(i)}\right]\left[1+\frac{wp(j)}{1-p(j)}\right]\right\}. \end{aligned}$$

We derive next $\bar{\varphi}_1(1) = 1$ so the singularity in (6) is removed.

$$\begin{aligned} g_w &= -\sum_{i=1}^{\infty} p(i)(1-p(i)) = -1/\alpha, \\ g_{ww} &= \sum_{i=1}^{\infty} 2p(i)^2(1-p(i)) = \frac{-2q(2q^2-q-1)}{(1+q)(q^2+q+1)}, \\ \bar{\varphi}_{1,w}(1) &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}[p(i)p(j) - p(i)^2p(j) - p(i)p(j)^2] = 1 - 2\sum_{i=1}^{\infty} p(i)^2 = \frac{3q-1}{1+q}. \end{aligned}$$

After all algebra, this finally leads to $\mathrm{VAR}(X(m)) \sim m\sigma^2$ where

$$\sigma^2 = \frac{-(q^3-q^2+2q-2)(1+q)}{4q^2(q^2+q+1)}.$$

We can check that our Markov chain is φ-mixing (see for instance Billingsley [4] p. 168) and we apply the Functional Central Limit Theorem([4], p.174, Thm. 20.1). We obtain the following result, where $B(t)$ is the standard Brownian Motion (BM).

Theorem 2.1

$$\frac{X([Nt]) - \bar{L}Nt}{\sigma\sqrt{N}} \Rightarrow B(t), N \to \infty, t \in [0,1],$$

Let us now condition on $X(m)=n$. M becomes a R.V. A realization of X for fixed n is given by Thm. 2.1, 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 $N=n/\bar{L}$ in the denominator of Thm.2.1 and we obtain the following result.

Theorem 2.2 *Conditioned on $X(m)=n$,*

$$\frac{X([mt])-\bar{L}mt}{\sigma\sqrt{n/\bar{L}}} \Rightarrow B(t)$$

Moreover, from [18] , the R.V. M:= number of runs is such that

$$M \sim \mathcal{N}(n\mu_1, n\sigma_1^2) \tag{7}$$

where

$$\begin{aligned}
\mathcal{N} &:= \text{a Gaussian(normal) R.V.},\\
\mu_1 &= 1/\bar{L} = \frac{2q}{1+q},\\
\sigma_1^2 &= \sigma^2/\bar{L}^3 = \frac{2(1-q)(q^2+2)q}{(1+q)^2(q^2+q+1)}.
\end{aligned}$$

This is actually an application of the Renewal Theorem.
Property (7) was already found in Grabner et al. [8].
Let us finally remark that we have taken here the opposite of our usual route: we have started from a Markov chain and constructed the corresponding polyomino, in contrast to our previous work on directed animals and compositions.

2.3 Runs length distribution

For large m (or n), the asymptotic PGF of the stationary distribution of the run length L is given by

$$\sum_{i=1}^{\infty}\pi(i)F(i,z) = \alpha z\sum_{i=1}^{\infty}\frac{p(i)(1-p(i))^2}{1-p(i)z}. \tag{8}$$

All moments of this distribution can be derived from (8) by setting $z=e^s$ and expanding wrt s. Of course, we recover $\mathbb{E}(L)=\bar{L}$ and $\mathbb{E}(L^2)=s^2$. Moreover, expanding (8) wrt z gives the asymptotic distribution $\bar{F}(l)$ of the run length L: $\bar{F}(l) := \Pr[L=l]$. For instance,

$$\begin{aligned}
\bar{F}(1) &= \frac{4p^2-9p+6}{2(p^2-3p+3)},\\
\bar{F}(2) &= \frac{p(2p^2-4p+3)(1-p)}{(p^2-3p+3)(p^2-2p+2)},\\
\bar{F}(3) &= \frac{p^2(4p^5-24p^4+61p^3-82p^2+60p-20)(p-1)}{2(p^2-2p+2)(p^2-3p+3)(p^4-5p^3+10p^2-10p+5)}.
\end{aligned}$$

The distribution is geometrically decreasing.
We have done a simulation of $U=2000$ sequences of $n=10000$ GEOM(0.75) R.V. (this simulation will be extensively used in Sec.3), leading to 8004396 runs. The observed and limiting run length distributions are given in Figure 1. The fit is quite good. A χ^2 test with 16 classes gives 10.45, which corresponds to a level of significance .79.

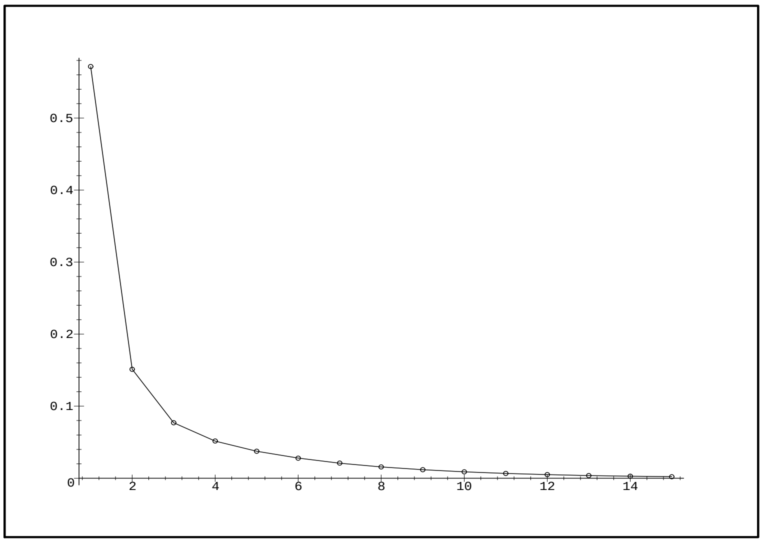

Figure 1: Run length(observed = circle, asymptotic = line)

3 Simulations and maximum run length distribution

In this section, we first present some trajectories simulations. We then study the hitting time to a length k run and the maximum run length distribution.

3.1 Realizations

We have simulated a typical trajectory for $X(i)$ for $n = 10000, p = 0.75$. This has given $M = 3997$ runs. The unnormalized trajectory is given in Figure 2, which shows a "filament silhouette". The normalized trajectory for $\frac{X(i)-\bar{L}i}{\sigma\sqrt{M}}$ and $\frac{X(i-1)-\bar{L}i}{\sigma\sqrt{M}}$, showing the run length is given in Figure 3. Of course, both trajectories are asymptotically equivalent for $n \to \infty$. A zoom on $i = [530 \dots 550]$ is given in Figure 4. We have also checked the asymptotic Gaussian property of M as given by (7). This is shown in Figure 5. A χ^2 test with 14 classes gives 12.87, which corresponds to a level of significance .46.

3.2 Hitting time and maximum run length distribution

In this section, all vectors and matrices will be written in bold. Let us define $T_k :=$ the time(counted in terms of number of GEOM R.V.) necessary to obtain a run of length k. We set $h_i := \mathbb{E}_i(T_k)$, (when possible, we drop the k for ease of notation). With (4) , we derive

$$h_i = [\bar{L}(i,k) + kp(i)^{k-1}] + \Lambda(i,k-1) \sum_{\substack{u=1 \\ u \neq i}}^{\infty} \Pi(i,u) h_u. \tag{9}$$

By (2), we observe that $\varphi(i,k) = \mathcal{O}(\varepsilon), \varepsilon := p^k, k \to \infty$, uniformly on i, and by standard properties of Markov chains(see Keilson [10], Aldous [1], Aldous and Brown [2], [3]), we know that the hitting time

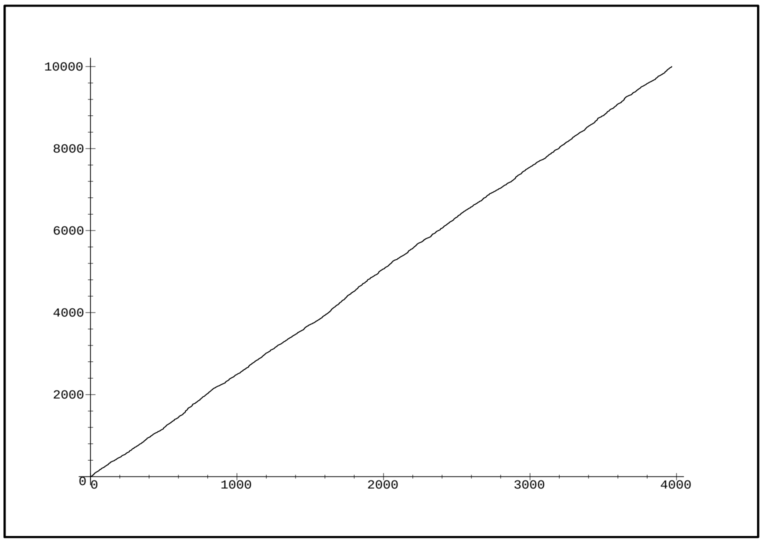

Figure 2: Unnormalized $X(\cdot)$

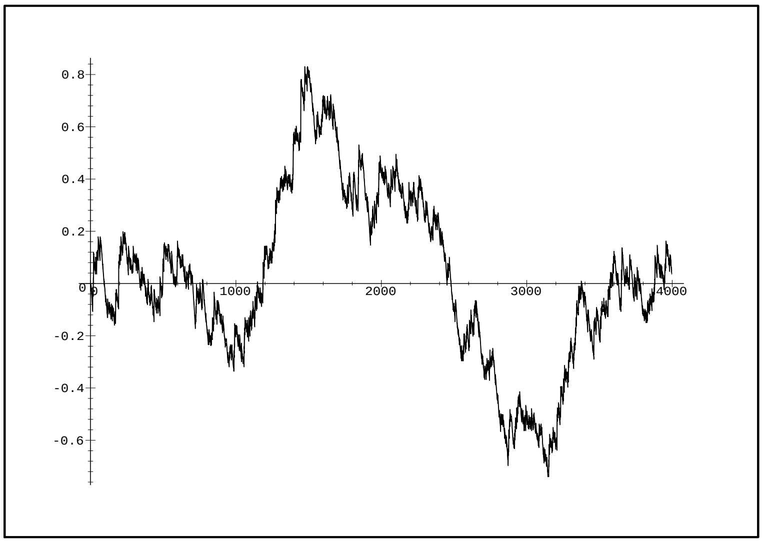

Figure 3: Normalized $X(\cdot)$, with run length

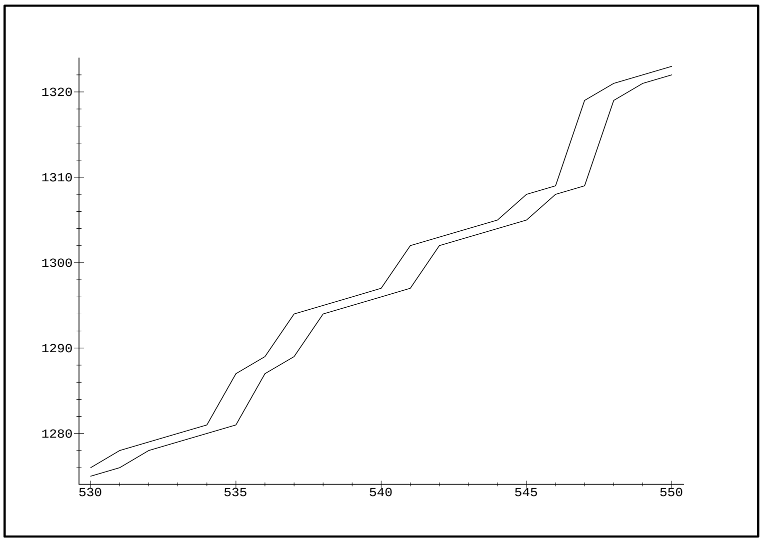

Figure 4: Unnormalized $X(\cdot)$, with run length(zoom)

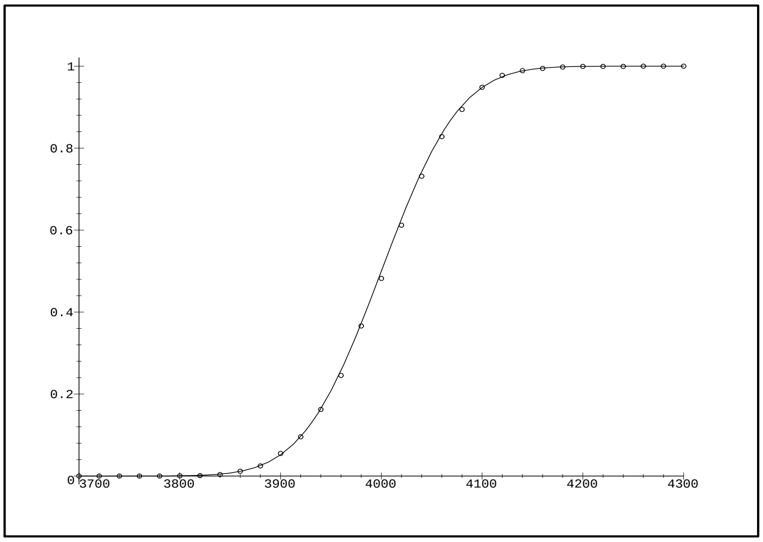

Figure 5: Observed and limiting M distribution

to k is such that:

$$\begin{aligned} h_i &= \frac{D(i)}{\varepsilon} + \psi(i) + \mathcal{O}(\varepsilon) \text{ (Actually a Laurent series exists for } \varepsilon \text{ sufficiently small)} \\ Pr_i[T_k \geq x] &\sim e^{-x/h_i}, x \to \infty. \end{aligned} \tag{10}$$

We will soon check that $\mathbf{D}$ is independent of i. First we note that

$$\bar{L}(i,k) + kp(i)^{k-1} = \bar{L}(i) + \mathcal{O}(\varepsilon).$$

Equ.(9) leads to

$$\begin{aligned} D(i) + \varepsilon\psi(i) &= \varepsilon\bar{L}(i) + \left[\sum_{\substack{u=1 \\ u \neq i}}^{\infty} \Pi(i,u)[D(u) + \varepsilon\psi(u)]\right]\left(1 - p(i)^{k-1}\right) + \mathcal{O}(\varepsilon^2) \\ &= \mathbf{\Pi D}(i) + \mathcal{O}(\varepsilon) \end{aligned} \tag{11}$$

Identification of ϵ powers in (11) leads to $[\mathbf{I} - \mathbf{\Pi}]\mathbf{D} = 0$, which confirms that $\mathbf{D}$ is independent of i. Now we premultiply (9) by $\boldsymbol{\pi}$. This leads to

$$\boldsymbol{\pi}\mathbf{h} = \bar{L} + \boldsymbol{\pi}\mathbf{h} - \sum_{i=1}^{\infty}\sum_{\substack{u=1 \\ u \neq i}}^{\infty} \alpha p(i)p(u)p(i)^{k-1}h_u + \mathcal{O}(\varepsilon)$$

or

$$\begin{aligned} 0 &= \bar{L} - \alpha\sum_{i=1}^{\infty} p(i)^k \sum_{u=1}^{\infty} p(u)h_u + \alpha\sum_{i=1}^{\infty} p(i)^{k+1}h_i + \mathcal{O}(\varepsilon) \\ &= \bar{L} - \alpha D p^k q/\varepsilon + \mathcal{O}(\varepsilon). \end{aligned}$$

Therefore, we obtain $D = 1/q$. We proceed now as in [16]. Let $\mathcal{M}(n) :=$ maximum run length based on n GEOM(p) R.V. . We know that

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

We start our asymptotic analysis by first using (10):

$$\Pr[\mathcal{M}(n) < k] \sim \exp[-\exp[\log n - k\log(p^{-1}) - \log D]], \quad n \to \infty$$

Set now $k := j+1$, $D_1 = p/D = pq$. We derive the following theorem

Theorem 3.1 *Let*

$$\eta := j - [\log n + \log D_1]/\log(p^{-1})$$

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

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

where $G_1(\eta) := \exp[-\exp[-\eta\log(p^{-1})]]$. Let

$$\psi_1(n) := \log(D_1 n)/\log(p^{-1})$$

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

$$\log \Pr[\mathcal{M}(n) \leq \lfloor \psi_1(n) \rfloor + k] e^{-\{\psi_1(n)\}\log(p^{-1})} \underset{n\to\infty}{\longrightarrow} - e^{-k\log(p^{-1})}$$

We also have

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

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

The asymptotic moments of $\mathcal{M}(n)$ are also periodic functions of $\psi_1(n)$. They can be written as Harmonic sums which are usually analyzed with Mellin transforms: see Flajolet et al [7]. The asymptotic non-periodic term in the moments of $\mathcal{M}(n)$ is given by the following result.

Theorem 3.2 *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*

$$\bar{\mathbb{E}}[\mathcal{M}(n) - \psi_1(n)]^i \sim \int_{-\infty}^{+\infty} \eta^i [G_1(\eta) - G_1(\eta - 1)] d\eta$$

Proof We need a rate of convergence property for Thm. 3.1. Indeed, the analysis of this type of Harmonic sums is detailed in Flajolet [6] , where the author uses bounds for tail estimates. In our case, there will be established as follows. From Keilson [10], we know that

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

for some cst C, uniformly on x and $h_i^{(2)} := \mathbb{E}_i(T_k^2)$. Now, by (14) and (18),

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

For $k\log(p^{-1})$ respectively given by $\log n - \log\log n$, $\log n$ and $2\log n$, the error bounds are $(\log n/n)^{1/4}$, $1/n^{1/4}$ and $1/n^{1/2}$, which are sufficient to establish our conclusions, following the lines of [6]. ■

It is well known that the extreme-value distribution function $e^{-e^{-x}}$ has mean γ and variance $\pi^2/6$. From this, we can for instance derive

$$\bar{\mathbb{E}}[\mathcal{M}(n)] \sim \psi_1(n) + \frac{1}{2} + \frac{\gamma}{\log(p^{-1})}. \tag{12}$$

The other periodic terms have very small amplitude (see Flajolet et al [7]).
From Hitczenko and Louchard [9], we also know that

$$\mathrm{VAR}[\mathcal{M}(n)] \sim \frac{\pi^2}{6\log(p^{-1})} + 1/12. \tag{13}$$

The dominant term ($\log n$) of (12) is already given in Grabner et al. [8]. The observed and limiting distribution function (DF) are given in Figure 6 (observed = circle, asymptotic = line). The limiting DF is given by $G_1[j - \psi_1(n)]$ and, due to the sensitivity of G_1 to the mean, we have taken $\psi_1(n) = \bar{\mathcal{M}}_0(n) - \frac{1}{2} - \frac{\gamma}{\log(p^{-1})}$ where $\bar{\mathcal{M}}_0(n)$ is the observed mean $= 28.8425\ldots$ The limiting mean $\bar{\mathbb{E}}(\mathcal{M}(n))$

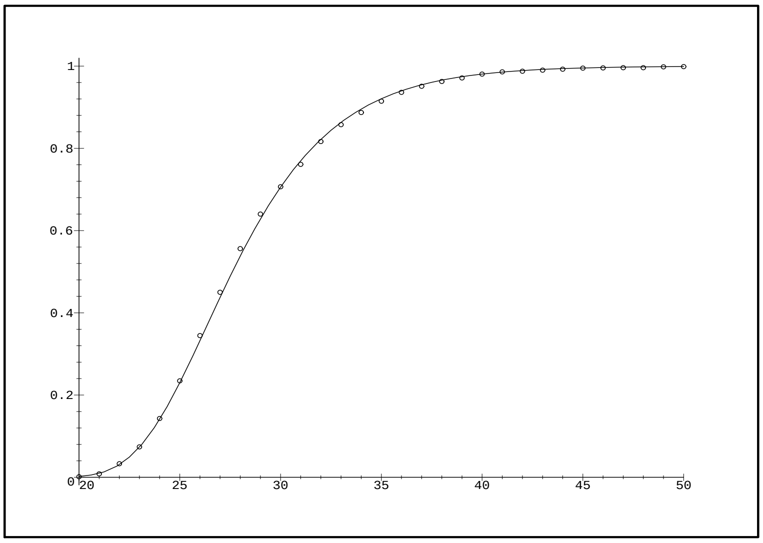

Figure 6: Observed and limiting $\mathcal{M}$ distribution function

is given by (12) and leads to 31.11270630... A χ^2 test with 14 classes gives 11.29, which corresponds to a level of significance .50. (We have here 12 degrees of freedom). Actually, $n = 10000$ is not large enough to entail the validity of our mean asymptotic expression. But larger simulations are time-consuming. So we have tested a sample of 100 sequences of 40000 GEOM(0.75) R.V. This has given an observed mean 33.03 that we must compare with the asymptotic value 33.52212715...After normalizing with (13), this leads to a normalized value 1.10 which corresponds to a level of significance 0.27 for a $\mathcal{N}(0,1)$.

Let us now consider the maximum run length of type u. We derive

$$
\begin{aligned}
h_i(u) &= \bar{L}(i) + \sum_{\substack{j=1 \\ j \neq i}}^{\infty} \Pi(i,j) h_j(u) \quad , i \neq u \\
h_u(u) &= [\bar{L}(u,k) + kp(u)^{k-1}] + \Lambda(u, k-1) \sum_{\substack{j=1 \\ j \neq u}}^{\infty} \Pi(u,j) h_j(u).
\end{aligned}
$$

Proceeding as above, with now $\varepsilon := p(u)^k$, we obtain

$$
h_i(u) \sim \frac{D(u)}{\varepsilon} + \mathcal{O}(1)
$$

with $D(u) = 1/q(u), q(u) := 1 - p(u)$. All previous results are easily adapted. For $u = 1$, the limiting distribution is given by Thm.3.1. Note that $p(u) \equiv p$ in this case. This means that the dominant term of the maximum run length is related to runs of type 1. For $u = 2$, we obtain an observed mean =5.2530... and a limiting mean =5.636922625.... The simulation is given in Figure 7 .A χ^2 test with 4 classes gives 1.067, which corresponds to a level of significance .59. (We have here 2 degrees of freedom). Again our sample of of 100 sequences of 40000 GEOM(0.75) R.V. leads to an observed mean 6.13 that we must compare with the asymptotic value 6.050994871...After normalizing with (13), this leads to a normalized value -0.96 which corresponds to a level of significance 0.33 for a $\mathcal{N}(0,1)$.

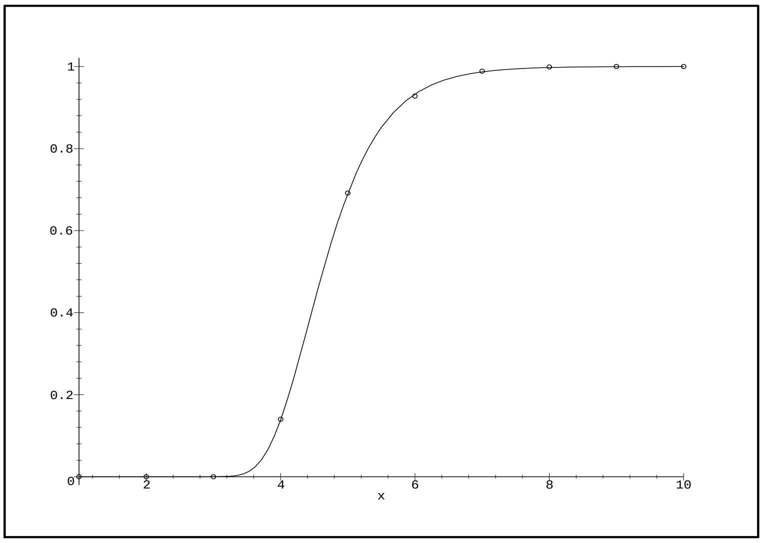

Figure 7: Observed and limiting $\mathcal{M}$ distribution function, u=2

4 Conclusion

Using a description of the runs either as a stochastic process or as a polyomino, we have obtained several asymptotic probability distributions. We intend to pursue this approach on other runs properties such as ascending runs (see, for instance, Prodinger [19]).

Acknowledgement
The pertinent comments of the referees led to substantial improvements in the presentation.

References

[1] D. Aldous. *Probability Approximations via the Poisson Clumping Heuristic.* Springer-Verlag, 1989.

[2] D. Aldous and M. Brown. Inequalities for rare events in time-reversible markov chains I. *Stochastic Inequalities, IMS*, 22:1–16, 1992.

[3] D. Aldous and M. Brown. Inequalities for rare events in time-reversible markov chains II. *Stochastic Processes and their Applications*, 44:15–25, 1993.

[4] P. Billingsley. *Convergence of Probability Measures.* Wiley, 1968.

[5] S. L. Campbell and C. D. Meyer. *Generalized Inverse of Linear Transformations.* Pitman, 1979.

[6] P. Flajolet. Approximate counting :a detailed analysis. *BIT*, 25:113–134, 1985.

[7] P. Flajolet, X. Gourdon, and P. Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144:3–58, 1995.

[8] P.J. Grabner, A. Knopfmacher, and H. Prodinger. Run statistics for geometrically distributed random variables. *LNCS*, 1776:457–462, 2000.

[9] P. Hitczenko and G. Louchard. Distinctness of compositions of an integer: a probabilistic analysis. In preparation.

[10] J. Keilson. *Markov Chain Models-Rarity and Exponentiality.* Springer-Verlag, 1979.

[11] J. G. Kemeny, J. L. Snell, and A. W. Knapp. *Denumerable Markov Chains.* Van Nostrand, 1966.

[12] P. Kirschenhofer and H. Prodinger. The path length of random skip lists. *Acta Informatica*, 31:775–792, 1994.

[13] P. Kirschenhofer, H. Prodinger, and W. Szpankowski. Analysis of a splitting process arising in probabilistic counting and other related algorithms. *Random Structures and Algorithms*, 9:379–401, 1996.

[14] D.E. Knuth. *The Art of Computer Programming.V.3.* Addison-Wesley, 1998.

[15] G. Louchard. Probalistic analysis of some (un)directed animals. *Theoretical Computer Science*, 159(1):65–79, 1996.

[16] G. Louchard. Probabilistic analysis of column-convex and directed diagonally-convex animals.II: Trajectories and shapes. *Random Structures and Algorithms*, 15:1–23, 1999.

[17] G. Louchard and H. Prodinger. Probabilistic analysis of Carlitz compositions. Submitted.

[18] G. Louchard and W. Szpankowski. Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithm. *IEEE Transactions on Information Theory*, 41:478–488, 1995.

[19] H. Prodinger. Combinatorics of geometrically distributed random variables: length of ascending runs. *LNCS*, 1776:473–482, 2000.

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

We must first be more precise in the analysis of h_i . Assume that $q < p$ (the analysis is similar if $q \geq p$). Set

$$h_i = \frac{D(i)}{\varepsilon} + \psi(i) + \varepsilon\chi(i) + \mathcal{O}(\varepsilon^2). \tag{14}$$

We note that

$$\bar{L}(i,k) + kp(i)^{k-1} = \bar{L}(i) + p^k q^{k(i-1)}/(q^{i-1} - q^i - 1).$$

We obtain

$$\begin{aligned} D(i) + \varepsilon\psi(i) + \varepsilon^2\chi(i) &= \varepsilon\bar{L}(i) + \varepsilon^2 q^{k(i-1)}/(q^{i-1} - q^i - 1) \\ &+ \quad [\mathbf{\Pi D}(i) + \varepsilon\mathbf{\Pi}\boldsymbol{\psi}(i) + \varepsilon^2\mathbf{\Pi}\boldsymbol{\chi}(i)]\left(1 - \varepsilon q^{(i-1)(k-1)}/p\right) + \mathcal{O}(\varepsilon^3). \end{aligned} \tag{15}$$

The ε term leads to

$$\psi(i) = \bar{L}(i) + \mathbf{\Pi}\boldsymbol{\psi}(i) - Dq^{(i-1)(k-1)}/p$$

or

$$[\mathbf{I} - \mathbf{\Pi}]\boldsymbol{\psi} = \boldsymbol{\delta} \tag{16}$$

where

$$\begin{aligned}\delta(1) &= \bar{L}(1) - 1/(pq)\\ \delta(i) &= \bar{L}(i), \quad i > 1.\end{aligned}$$

First we premultiply (16) by $\boldsymbol{\pi}$. This leads to

$$\alpha = \bar{L} = D\pi(1)/p \quad \text{hence} \quad D = 1/q$$

which confirms to the analysis of Sec.3.2. Next (16) leads to

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

and $\mathbf{M} := \sum_{n\geq 0}(\mathbf{\Pi}^n - \mathbf{1}\times\boldsymbol{\pi})$. $\mathbf{M}$ is the Drazin inverse of $\mathbf{I}-\mathbf{\Pi}$. We refer to Campbell and Meyer [5] for a detailed definition and analysis of the Drazin inverse. We have $\mathbf{M} = \mathbf{Z} - \mathbf{1}\times\boldsymbol{\pi}$ where $\mathbf{Z} := [I - \mathbf{\Pi} + \mathbf{1}\times\boldsymbol{\pi}]^{-1} = \sum_{n\geq 0}[\mathbf{\Pi} - \mathbf{1}\times\boldsymbol{\pi}]^n$ is the potential used in Kemeny, Snell and Knapp [11]. First, by (22)(see the next Appendix section), we know that $\mathbf{M}\boldsymbol{\delta}$ is finite . Next, to derive γ_1, we premultiply (15) by $\boldsymbol{\pi}$. This leads to

$$D+\varepsilon\boldsymbol{\pi}\boldsymbol{\psi}+\varepsilon^2\boldsymbol{\pi}\boldsymbol{\chi} = \varepsilon\bar{L}-\varepsilon^2\pi(1)/q+D+\varepsilon\boldsymbol{\pi}\boldsymbol{\psi}+\varepsilon^2\boldsymbol{\pi}\boldsymbol{\chi}-\sum_i\sum_{u\neq i}\alpha p(i)p(u)[D+\varepsilon\psi(u)+\varepsilon^2\chi(u)]p(i)^{k-1}+\mathcal{O}(\varepsilon^3).$$

The last term can be written as

$$-\alpha\frac{p^k}{1-q^k}[D+\varepsilon\sum_u p(u)\psi(u)] + \alpha\frac{p^{k+1}}{1-q^{k+1}}D + \alpha\varepsilon\sum_i p(i)^{k+1}\psi(i) + \mathcal{O}(\varepsilon^3).$$

The ε term leads of course to $D = 1/q$. The ε^2 term leads to

$$0 = -\pi(1)/q - \alpha\sum_u p(u)\psi(u) + \alpha p^2\psi(1)$$

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 obtain, after some algebra,

$$h_i^{(2)} = s^2(i) + \varepsilon q^{k(i-1)}G(k,i) + 2\bar{L}(i,k)\sum_{\substack{u=1\\u\neq i}}^{\infty}\Pi(i,u)h_u + \Lambda(i,k-1)\sum_{\substack{u=1\\u\neq i}}^{\infty}\Pi(i,u)h_u^{(2)} \tag{17}$$

where $s^2(i)$ is given by (3) and $G(k,i)$ is a (complicated) polynomial in k, q^i. When $i=1$, $G(k,1)$ is linear in k. We set

$$h_i^{(2)} = \frac{D^{(2)}(i)}{\varepsilon^2} + \frac{\psi^{(2)}(i)}{\varepsilon}\chi^{(2)}(i) + \mathcal{O}(\varepsilon). \tag{18}$$

Equ.(17) leads to

$$\begin{aligned}&D^{(2)}(i) + \varepsilon\psi^{(2)}(i) + \varepsilon^2\chi^{(2)}(i)\\ =\;& \varepsilon^2 s^2(i) + 2[\varepsilon\bar{L}(i) + \varepsilon^2 q^{k(i-1)}/(q^{i-1}-q^i-1) - k\varepsilon^2 q^{(i-1)(k-1)}/p][\mathbf{\Pi}\mathbf{D}(i) + \varepsilon\mathbf{\Pi}\boldsymbol{\psi}(i)]\\ +\;& [\mathbf{\Pi}\mathbf{D}^{(\mathbf{2})}(i) + \varepsilon\mathbf{\Pi}\boldsymbol{\psi}^{(\mathbf{2})}(i) + \varepsilon^2\mathbf{\Pi}\boldsymbol{\chi}^{(\mathbf{2})}(i)]\left(1 - \varepsilon q^{(i-1)(k-1)}/p\right) + \mathcal{O}(\varepsilon^3).\end{aligned} \tag{19}$$

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

$$\psi^{(2)}(i)=2\bar{L}(i)D+\mathbf{\Pi}\boldsymbol{\psi}^{(2)}(i)-D^{(2)}q^{(i-1)(k-1)}/p$$

or

$$[\mathbf{I}-\mathbf{\Pi}]\boldsymbol{\psi}^{(2)}=\boldsymbol{\delta}^{(2)} \tag{20}$$

where

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

First we premultiply (20) by $\boldsymbol{\pi}$. This leads to

$$0=2\bar{L}D-D^{(2)}\pi(1)/p \quad \text{hence} \quad D^{(2)}=2/q^2\equiv 2D^2 \quad \text{as it should.}$$

Next (20) leads to

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

First, by (22) , we know that $\mathbf{M}\boldsymbol{\delta}^{(2)}$ is finite . Next, to derive γ_2, we premultiply (19) by $\boldsymbol{\pi}$. This leads to

$$\begin{aligned}&D^{(2)}+\varepsilon\boldsymbol{\pi}\boldsymbol{\psi}^{(2)}+\varepsilon^2\boldsymbol{\pi}\boldsymbol{\chi}^{(2)}\\ &= \varepsilon^2 s^2+2\varepsilon\bar{L}D+2\varepsilon^2\sum_i \pi(i)\bar{L}(i)\mathbf{\Pi}\boldsymbol{\psi}(i)+2\varepsilon^2\pi(1)(-1/q-k/p)D+D^{(2)}+\varepsilon\boldsymbol{\pi}\boldsymbol{\psi}^{(2)}+\varepsilon^2\boldsymbol{\pi}\boldsymbol{\chi}^{(2)}\\ &- \sum_i\sum_{u\neq i}\alpha p(i)p(u)[D^{(2)}+\varepsilon\psi^{(2)}(u)]p(i)^{k-1}+\mathcal{O}(\varepsilon^3).\end{aligned}$$

The last term can be written as

$$-\alpha\frac{p^k}{1-q^k}[D^{(2)}+\varepsilon\sum_u p(u)\psi^{(2)}(u)]+\alpha\frac{p^{k+1}}{1-q^{k+1}}D^{(2)}+\alpha\varepsilon\sum_i p(i)^{k+1}\psi^{(2)}(i)+\mathcal{O}(\varepsilon^3).$$

The ε term leads of course to $D^{(2)}=1/q^2$. The ε^2 term leads to

$$0=s^2+2\sum_i \pi(i)\bar{L}(i)\mathbf{\Pi}\boldsymbol{\psi}(i)+2\pi(1)(-1/q-k/p)D-\alpha\sum_u p(u)\psi^{(2)}(u)+\alpha p^2\psi^{(2)}(1)$$

which allows the determination of γ_2.

B Appendix : An explicit expression for M

We proceed as in Sec.2.2. We must compute $\mathbf{M}:=\sum_{n\geq 0}(\mathbf{\Pi}^n-\mathbf{1}\times\boldsymbol{\pi})$.
We see that $\mathbf{S}:=\sum_{n\geq 1}(\mathbf{\Pi}^n-\mathbf{1}\times\boldsymbol{\pi})$ can be written as

$$S(j,k)=\lim_{w\to 1}\left[\sum_{n\geq 1} w^{n+1}\Pi^n(j,k)+\frac{w^2}{1-w}\alpha p(j)(1-p(j))\right].$$

But

$$\Pi(j,k)=\frac{1}{p(i)}\frac{p(i)}{1-p(i)}[i\neq j]\frac{p(j)}{1-p(j)}(1-p(j)),$$

similarly,

$$\Pi^2(j,k) = \frac{1}{p(i)} \frac{p(i)}{1-p(i)} \sum_{l \neq i,j} \frac{p(l)}{1-p(l))} \frac{p(j)}{1-p(j)} (1-p(j)).$$

We obtain

$$S(j,k) = \lim_{w \to 1} \left[\frac{1}{p(i)} [\theta^j][\phi(w,\theta,z^*|i) - w\theta^i x^{*i} z^*/(1-x^{*i}z^*)](1-p(j)) - \frac{w^2}{1-w} \alpha p(j)(1-p(j)) \right]$$

or

$$\begin{aligned} S &= \lim_{w \to 1} \left[\frac{1}{p(i)} [\theta^j] \left\{ \frac{A_2(w,\theta,x^*,z^*|i) - w\theta^i x^{*i} z^*/(1-x^{*i}z^*)}{w} \right. \right. \\ &+ \left. \left. \frac{A_1(w,\theta,x^*,z^*) A_2(w,1,x^*,z^*|i)}{wg(w,x^*,z^*)} \right\} (1-p(j)) - \frac{w}{1-w} \alpha p(j)(1-p(j)) \right] \end{aligned}$$

This leads to (we set $w = 1 - \varepsilon$)

$$\begin{aligned} S = \lim_{\epsilon \to 0} &\left[-\frac{-\alpha p(j)(1-p(j))}{\epsilon} + \alpha p(j)(1-p(j)) - p(i)[i=j] \right. \\ &+ \left. \frac{1}{p(i)} \left[-\frac{\bar{\varphi}_1(1)}{g_w \epsilon} + \frac{\bar{\varphi}_{1,w}(1)}{g_w} - \frac{\bar{\varphi}_1(1) g_{ww}}{2g_w^2} \right] (1-p(j)) \right] \end{aligned} \tag{21}$$

where

$$\varphi_1(w) = w^2 \frac{p(i)}{1-p(i)} \frac{p(j)}{1-p(j)} \bigg/ \left\{ w \left[1 + \frac{wp(i)}{1-p(i)} \right] \left[1 + \frac{wp(j)}{1-p(j)} \right] \right\}.$$

We derive next $\bar{\varphi}_1(1) = p(i)p(j)$ so the singularity in (21) is removed.

$$\bar{\varphi}_{1,w}(1) = [p(i)p(j) - p(i)^2 p(j) - p(i)p(j)^2].$$

And finally,

$$\begin{aligned} M(i,j) &= S(i,j) + [i=j] - \pi(j) \\ &= [i=j] - p(i)[i=j] \\ &+ \frac{1}{p(i)} [-\alpha[p(i)p(j) - p(i)^2 p(j) - p(i)p(j)^2] - p(i)p(j) g_{ww} \alpha^2/2](1-p(j)). \end{aligned} \tag{22}$$