Distinctness of compositions of an integer: A Probabilistic Analysis

Paweł Hitczenko* and Guy Louchard †

April 18, 2001

Abstract

Compositions of integers are used as theoretical models for many applications. The degree of distinctness of a composition is a natural and important parameter. In this paper, we use as measure of distinctness the number of distinct parts (or components). We investigate, from a probabilistic point of view, the first empty part, the maximum part size and the distribution of the number of distinct part sizes. We obtain asymptotically, for the classical composition of an integer, the moments and an expression for a continuous distribution F, the (discrete) distribution of the number of distinct part sizes being computable from F. We next analyze another composition: the Carlitz one, where two successive parts are different. We use tools such as analytical depoissonization, Mellin transforms, Markov chain potential theory, limiting hitting times, singularity analysis and perturbation analysis.

*Department of Mathematics and Computer Science, Drexel University, Philadelphia, PA 19104, email: phitczen@mcs.drexel.edu

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

1 Introduction

Compositions of integers are used as theoretical models for many applications. The degree of distinctness of a composition is a natural and important parameter. Many references and applications can be found in Hwang and Yeh, [21] which attracted our interest into this fascinating topic. We consider the composition of an integer N into k parts, $(\gamma_1, \ldots, \gamma_k)$ i.e. $N = \sum_i^k \gamma_i, \quad \gamma_i$: integer > 0. We define the indicator variable $I_i :=$ [value i appears among these k values]. Considering all compositions as equiprobable, we are interested in stochastic properties of the distinctness measured by $\mathcal{D}_N := \sum_i I_i$. In this paper we consider the asymptotic properties of the distribution function of the number of distinct part sizes in a randomly chosen composition of an integer N. Investigation of random compositions or partitions from the probabilistic perspective originated over six decades ago with a paper by Erdös and Lehner [9] who studied the limiting distribution of the total number of parts in a random partition. Since then several other quantities have been studied. One of them is the number of distinct part sizes. For partitions, Wilf [40] found an asymptotic formula for the expected number of distinct part sizes. Subsequently, Goh and Schmutz [18] established the central limit theorem for the number of distinct part sizes in a randomly chosen partition.

For compositions, the question has not been settled. As far as the expected value of the number of distinct part sizes in a random composition, Knopfmacher and Mays [29] obtained the generating function which could presumably be analyzed to yield the asymptotic behavior. Hwang and Yeh [21] used generating function approach and, among other things, derived explicit formulas for the asymptotics of this expectation. The same result was independently, but later obtained in [20] using entirely different probabilistic approach that was developed in [19]. Hwang and Yeh raised the question about the asymptotics for the distribution function of the number of distinct part sizes. In this paper we couple the probabilistic approach of [19] with generating function method and poissonization techniques of [10], [23], [24] to address that question. We obtain asymptotically, for the classical composition of an integer, the moments and an expression for a continuous distribution F, the (discrete) distribution of $\mathcal{D}_N$ being computable from F. We also investigate two related quantities, namely, first empty part, $\mathcal{E}_N$, and the maximum part size $\mathcal{M}_N$.

Furthermore, we analyze another composition: the Carlitz one, where two successive parts are different. Some aspects of this composition have already been considered in [17], [30] and [37]. In addition to poissonization/depoissonization, Markov chain potential theory, and limiting hitting times , we use such analytical tools as Mellin transform, singularity analysis, saddle point method and perturbation analysis.

The paper is organized as follows: in Sec.2, we consider the classical composition. In Sec.3 we give some asymptotic distributions for $\mathcal{D}_N$, $\mathcal{E}_N$, $\mathcal{M}_N$. Sec.4 is devoted to the Carlitz composition. Sec.5 concludes the paper. An Appendix provides some technical tools from potential theory and Drazin inverse, which are necessary in Sec.4.

2 Classical Compositions

In this section, we first present a formalization of a relationship between random compositions of integers and sequences of i.i.d. geometic random variables (r.v.s, for short) with parameter $1/2$. We then study the number of distinct part sizes, the maximum part size, and the first empty part in a random composition.

2.1 Representation of random compositions

The following representation of a composition that can be found in [4] is of crucial importance: there is a one-to-one correspondence between compositions of N and strings of black and white dots of length N with the following provisions:

(i) the last dot is always black

(ii) each of the remaining $N-1$ dots is black or white; part sizes in a composition are “waiting times” for the first, second,..., and kth appearance of a black dot.

Thus, for example, the string

$$\underbrace{\bullet}_{1}\ \ \underbrace{\circ\ \ \circ\ \ \bullet}_{3}\ \ \underbrace{\circ\ \ \bullet}_{2}\ \ \underbrace{\bullet}_{1}\ \ \underbrace{\bullet}_{1}\ \ \underbrace{\circ\ \ \bullet}_{2}\ \ \underbrace{\circ\ \ \bullet}_{2}$$

represents the composition of 12 into parts $(1,3,2,1,1,2,2)$. Considering random composition corresponds to having black and white dots on each of the first $N-1$ positions distributed like i.i.d. Bernoulli r.v.s. Waiting times have known distribution and after making correction for the last part we find that a random composition of N is equidistributed with

$$(\Gamma_1, \Gamma_2, \ldots, \Gamma_{\tau-1}, N - \sum_{j=1}^{\tau-1} \Gamma_j),$$

where $\Gamma_1, \Gamma_2 \ldots$ are i.i.d. geometric r.v.s with parameter $1/2$, GEOM$(1/2)$. That is

$$\mathsf{Pr}(\Gamma_1 = j) = \frac{1}{2^j}, \qquad j = 1, 2 \ldots,$$

and τ is defined by

$$\tau = \inf\{k \geq 1 : \ \Gamma_1 + \Gamma_2 + \ldots + \Gamma_k \geq N\}.$$

Since τ, being a $1 + \mathrm{Bin}(N-1, 1/2)$ r.v., is tightly concentrated around its expected value, it follows that the distribution of a random composition is close to that of

$$(\Gamma_1, \Gamma_2, \ldots, \Gamma_{\lfloor (N+1)/2 \rfloor}),$$

where $\lfloor x \rfloor$ is the integer part of x. We refer the reader to [19] or [20] for more details and precise statements. For the purpose of this paper it will be enough to record the following fact. Let $n = \lfloor (N+1)/2 \rfloor$ and for a composition $\kappa = (\gamma_1, \ldots, \gamma_k)$ let

$$\mathcal{D}_N(\kappa) = 1 + \sum_{i=2}^{k} I_{\{\gamma_i \neq \gamma_j,\ j=1,\ldots,i-1\}}$$

denote the number of distinct part sizes in a composition κ and let

$$D_n = 1 + \sum_{i=2}^{n} I_{\{\Gamma_i \neq \Gamma_j,\ j=1,\ldots,i-1\}}$$

be the number of distinct values in a sample of n i.i.d. GEOM(1/2) r.vs. Then by repeating the argument of [19, section 4] we infer that

$$|\mathsf{Pr}(\mathcal{D}_N \le t) - \mathsf{Pr}(D_n \le t)| = \mathcal{O}\left(\sqrt{\frac{\log N}{N}}\right).$$

Therefore it suffices to approximate the distribution function (DF) of the number of distinct values in the sequence of n i.i.d. geometric r.v.s. For this reason, from now on we set $n = \lfloor (N+1)/2 \rfloor$ and we consider n i.i.d. GEOM(1/2) r.v.s. We also note that τ represents the number of parts and thus the asymptotic distribution of that number is $\mathcal{N}(\frac{N}{2}, \frac{N}{4})$, where $\mathcal{N}$ denotes the Gaussian (normal) r.v. In general, just as we have done with the number of distinct part sizes, we will denote quantities of interest for compositions by script letters and the corresponding quantities for geometric variables will be denoted by the same letters from the roman alphabet (ocassionally we will drop the subscripts N and n, respectively). As for other notational conventions, log will denote $\log_2$, $L = \ln 2$ and β's (with or without subscripts) will be used to denote periodic functions of $\log n$, with mean 0, period 1 and with small (of order no more than 10^{-6}) amplitude. Actually, these functions depend on the fractional part of $\log n$: $\{\log n\}$.

2.2 Recurrence

To obtain a recurrence relation we condition on the number of Γ_j's that are equal to 1. Letting $D_0 \equiv 0$, by the law of total probability we find that

$$\begin{aligned}
\mathsf{Pr}(D_n = k) &= \sum_{j=0}^{n} \mathsf{Pr}(\{D_n = k\} \cap \{\sum_{\ell=1}^{n} I_{\Gamma_\ell = 1} = n - j\}) \\
&= \sum_{j=0}^{n} \mathsf{Pr}(D_n = k \Big| \sum_{\ell=1}^{n} I_{\Gamma_\ell=1} = n-j) \binom{n}{n-j} \frac{1}{2^n} \\
&= \sum_{j=0}^{n-1} \mathsf{Pr}(D_n = k \Big| \sum_{\ell=1}^{n} I_{\Gamma_\ell=1} = n-j) \binom{n}{n-j} \frac{1}{2^n}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2^n}\mathsf{Pr}(D_n = k \Big| \sum_{\ell=1}^{n} I_{\Gamma_\ell=1} = 0) \\
&= \sum_{j=0}^{n-1} \mathsf{Pr}(D_j = k-1)\binom{n}{n-j}\frac{1}{2^n} + \frac{\mathsf{Pr}(D_n = k)}{2^n},
\end{aligned}
$$

where, in the last line we have used

$$
\mathsf{Pr}(D_n = k \Big| \sum_{\ell=1}^{n} I_{\Gamma_\ell=1} = 0) = \mathsf{Pr}(D_j = k \Big| \forall\ell\ \Gamma_\ell \geq 2) = \mathsf{Pr}(D_n = k),
$$

which follows from the fact that given the event $\{\forall\ell\ \Gamma_\ell \geq 2\}$, (Γ_j) are i.i.d. random variables distributed like $1+\Gamma$. Therefore, the generating function of the probability distribution function of D_n is

$$
\begin{aligned}
G_n(u) &= \sum_{k\geq 0} \mathsf{Pr}(D_n = k)u^k = \\
&= \sum_{k\geq 0}\left(\sum_{j=0}^{n-1} \mathsf{Pr}(D_j = k-1)\binom{n}{j}\frac{1}{2^n} + \frac{\mathsf{Pr}(D_n = k)}{2^n}\right)u^k \\
&= \sum_{j=1}^{n-1}\binom{n}{j}\frac{1}{2^n}\sum_{k\geq 0}\mathsf{Pr}(D_j = k-1)u^k + \frac{1}{2^n}\sum_{k\geq 0}\mathsf{Pr}(D_n = k)u^k \\
&= \sum_{j=1}^{n-1}\binom{n}{j}\frac{1}{2^n}uG_j(u) + \frac{G_n(u)}{2}^n.
\end{aligned}
$$

Since $G_0(u) \equiv 1$, we obtain the following recurrence for G_n:

$$
G_n(u) = \begin{cases} 1 & \text{if } n = 0, \\ u\sum_{j=0}^{n-1}\binom{n}{j}\frac{G_j(u)}{2^n} + \frac{G_n(u)}{2^n} & \text{if } n \geq 1. \end{cases}
$$

Recurrences like that are very common in the analysis of certain algorithms and we will follow techniques developed for the purpose of studying them. We first consider the poissonized version:

$$
\begin{aligned}
G(z,u) &= \sum_{n=0}^{\infty} G_n(u)\frac{z^n e^{-z}}{n!} \\
&= G_0(u)e^{-z} + \sum_{n=1}^{\infty}\frac{z^n e^{-z}}{n!}\left\{u\sum_{j=0}^{n-1}\binom{n}{j}\frac{G_j(u)}{2^n} + \frac{G_n(u)}{2^n}\right\} \\
&= e^{-z} + \sum_{n=1}^{\infty}\frac{e^{-z}}{n!}\left(\frac{z}{2}\right)^n G_n(u) + u\sum_{n=1}^{\infty}\sum_{j=0}^{n-1}\frac{z^n e^{-z} G_j(u)}{2^n j!(n-j)!}
\end{aligned}
$$

$$\begin{aligned}
&= e^{-z} + e^{-z/2}\sum_{n=1}^{\infty}\frac{(z/2)^n e^{-z/2}}{n!}G_n(u) + u\sum_{j=0}^{\infty}\sum_{n=j+1}^{\infty}\frac{z^n e^{-z}G_j(u)}{2^n j!(n-j)!} \\
&= e^{-z} + e^{-z/2}\left\{G(z/2,u) - e^{-z/2}\right\} \\
&\qquad + u\sum_{j=0}^{\infty}\frac{G_j(u)}{j!}\left(\frac{z}{2}\right)^j e^{-z}\sum_{n=j+1}^{\infty}\frac{(z/2)^{n-j}}{(n-j)!} \\
&= e^{-z/2}G(z/2,u) + u\sum_{j=0}^{\infty}\frac{G_j(u)}{j!}\left(\frac{z}{2}\right)^j e^{-z}\sum_{m=1}^{\infty}\frac{(z/2)^m}{m!} \\
&= e^{-z/2}G(z/2,u) + u\sum_{j=0}^{\infty}\frac{G_j(u)}{j!}\left(\frac{z}{2}\right)^j e^{-z}\left\{e^{z/2}-1\right\} \\
&= e^{-z/2}G(z/2,u) + uG(z/2,u) - ue^{-z/2}G(z/2,u) \\
&= G(z/2,u)\left\{e^{-z/2}(1-u) + u\right\}
\end{aligned}$$

Hence, the function $H(z,u) = \frac{G(z,u)}{1-u}$ satisfies the same identity

$$H(z,u) = H(z/2,u)\left\{e^{-z/2} + u(1-e^{-z/2})\right\}, \tag{1}$$

which, by iteration, gives

$$H(z,u) = \frac{1}{1-u}\prod_{j=1}^{\infty}\left(e^{-z/2^j} + u(1-e^{-z/2^j})\right) \tag{2}$$

Observe that if $z > 0$ then $H(z,\ \cdot\ )$ is the generating function of the sequence $\mathsf{Pr}(X_z \le k)$, where $X_z = \sum_{j\ge 1} X_{z,j}$ and $(X_{z,j})$ are independent random variables satisfying

$$\mathsf{Pr}(X_{z,j} = 0) = e^{-z/2^j} \quad \text{and} \quad \mathsf{Pr}(X_{z,j} = 1) = 1 - e^{-z/2^j}. \tag{3}$$

We will now depoissonize that result by appealing to Jacquet and Szpankowski [23] (see also [24]). Let

$$H_k(z) = \sum_n \mathsf{Pr}(D_n \le k)\frac{z^n e^{-z}}{n!},$$

so that

$$H(z,u) = \sum_{k=0}^{\infty} H_k(z)u^k.$$

Comparison of coefficients of the powers of u on both sides of (1) and (2) yields, respectively,

$$H_k(z) = e^{-z/2}H_k(z/2) + (1-e^{z/2})H_{k-1}(z/2), \quad k \ge 0, \quad (H_{-1}(z) = 0), \tag{4}$$

and

$$H_k(z) = \sum_{m=0}^{k} \sum_{\substack{J \subset \mathbf{N} \\ |J|=m}} \prod_{j \in J} (1 - e^{-z/2^j}) \prod_{j \notin J} e^{-z/2^j}. \tag{5}$$

We will check that the functions $H_k(z)$ satisfy the conditions (I) and (O) of [23, Corollary 1]: $\exists\, 0 < \theta < \pi/2$, such that for a linear cone $\mathcal{S}_\theta = \{z : |\arg(z)| < \theta\}$ there exist $A, B, R > 0$, β and $\alpha < 1$ for which the following two conditions hold uniformly in $k \geq 0$:

(I) For $z \in \mathcal{S}_\theta$, $\quad |z| > R \Longrightarrow |H_k(z)| \leq B|z|^\beta$,

(O) For $z \notin \mathcal{S}_\theta$, $\quad |z| > R \Longrightarrow |H_k(z)e^z| \leq Ae^{\alpha|z|}$.

To verify (O) note that

$$|e^z H_k(z)| \leq \sum_{n=0}^{\infty} \frac{|z|^n}{n!} = e^{|z|},$$

so that (4) implies that

$$\begin{aligned} |e^z H_k(z)| &\leq |e^{z/2} H_k(z/2)| + |e^{z/2}(1 - e^{-z/2})||e^{z/2} H_{k-1}(z/2)| \\ &\leq e^{|z|/2}\left(1 + e^{\Re(z)/2} + 1\right) \leq 3e^{\alpha|z|}, \end{aligned}$$

for some $1/2 < \alpha < 1$ and large $|z|$.

We now show that (I) holds with $\beta = 0$. Let $z = x + iy$ be in $\mathcal{S}_\theta$. Then, by elementary manipulations

$$\begin{aligned} |1 - e^{-z/2^j}| &\leq 1 - e^{-x/2^j} + \frac{1 - \cos(y/2^j)}{\exp(x/2^j) - 1} \leq 1 - e^{-x/2^j} + \frac{y^2/2^{2j}}{2(\exp(x/2^j) - 1)} \\ &\leq 1 - e^{-x/2^j} + \frac{\tan^2\theta}{2} \frac{x^2/2^{2j}}{\exp(x/2^j) - 1}, \end{aligned}$$

and it follows from (5) that

$$\begin{aligned} |H_k(z)| &\leq \sum_{m=0}^{k} \sum_{\substack{J \subset \mathbf{N} \\ |J|=m}} \prod_{j \in J} |1 - e^{-z/2^j}| \prod_{j \notin J} |e^{-z/2^j}| \\ &\leq \sum_{m=0}^{\infty} \sum_{\substack{J \subset \mathbf{N} \\ |J|=m}} \prod_{j \in J} \left(1 - e^{-x/2^j} + \frac{\tan^2\theta}{2} \frac{x^2/2^{2j}}{\exp(x/2^j) - 1}\right) \prod_{j \notin J} e^{-x/2^j} \\ &= \prod_{j=1}^{\infty} \left(1 + \frac{\tan^2\theta}{2} \frac{x^2/2^{2j}}{\exp(x/2^j) - 1}\right) \\ &\leq \exp\left\{\frac{\tan^2\theta}{2} \sum_{j=1}^{\infty} \frac{(x/2^j)^2}{\exp(x/2^j) - 1}\right\}, \end{aligned}$$

which is bounded independently of x and k since, with $f(t) = \dfrac{x^2/2^{2t}}{\exp(x/2^t)-1}$, we have

$$\begin{aligned}
\sum_{j=1}^{\infty} \frac{(x/2^j)^2}{\exp(x/2^j)-1} &= \sum_{j=1}^{\infty} f(j) \le \sup_j f(j) + \int_0^{\infty} f(t)dt \\
&\le 1 + \frac{1}{\ln 2}\int_0^x \frac{u}{e^u-1}du \le 1 + \frac{\pi^2}{6\ln 2}.
\end{aligned}$$

It follows from Corollary 1 of [23] that

$$\mathsf{Pr}(D_n \le k) = H_k(n) + \mathcal{O}(\frac{1}{n}), \tag{6}$$

uniformly in $k \ge 0$. Hence, (5) implies that the asymptotics of the distribution of D_n (and thus also of $\mathcal{D}_N$) is the same as that of

$$X_n = \sum_{j=1}^{\infty} X_{n,j}, \tag{7}$$

where $(X_{n,j})$ are independent random variables given by (3). Thus, we will turn our attention to the sequence (X_n).

2.3 Properties of the distribution

W begin by noting that the representation (7) gives, by Mellin transform (see below), the asymptotic value of the expected number of D_n

$$\mathsf{E}D_n \sim \sum_{i=1}^{\infty} \mathsf{Pr}(X_{n,i}=1) = \sum_{i=1}^{\infty}\left\{1-e^{-n/2^i}\right\} \sim \log n - \frac{1}{2} + \frac{\gamma}{L} + \beta_1(\log n) + \mathcal{O}(1/n).$$

As $n = \lfloor (N+1)/2 \rfloor$, in order to find $\mathsf{E}\mathcal{D}_N$ we must replace $\log n$ by $\log N - 1$. Of course the mean conforms to Hwang and Yeh result [21]. Similarly, for the variance we obtain

$$\begin{aligned}
\mathsf{var}(D_n) &\sim \mathsf{var}(X_n) = \sum_{j=1}^{\infty} \mathsf{var}(X_{n,j}) = \sum_{j=1}^{\infty}(1-e^{-n/2^j})e^{-n/2^j} \\
&= \sum_{j=1}^{j_n}(1-e^{-n/2^j})e^{-n/2^j} + \sum_{j>j_n}(1-e^{-n/2^j})e^{-n/2^j} \\
&= \sum_{j=1}^{j_n}(e^{-n/2^j} - e^{-n/2^{j-1}}) + \mathcal{O}(\sum_{j>j_n}\frac{n}{2^j}) \\
&= (e^{-n/2^{j_n}} - e^{-n}) + \mathcal{O}(\frac{n}{2^{j_n}}) \to 1,
\end{aligned}$$

provided j_n is chosen so that $n/2^{j_n} \to 0$ as $n \to \infty$. Thus, while there is a periodic component in the limiting distribution, the variance has no periodicity. The same phenomena appears in the mean of adaptive sampling ([35]).

We now turn to the question of the weak convergence of the sequence $D_n - \lfloor \log n \rfloor$. As we will see, due to contribution of the fractional part of $\log n$, this sequence does not have a weak limit. It does, however, converge in distribution along subsequences n_m for which the fractional part of $\log n_m$ is about constant. More specifically, for a fixed $n_0 \geq 1$, let $x = \{\log n_0\}$ so that $n_0 = 2^{\lfloor \log n_0 \rfloor + x}$ and consider a subsequence (n_m) defined by $n_m = 2^{\lfloor \log n_0 \rfloor + m + x}$. Note that the array $(X_{n,j})$ has the following property

$$X_{2n,j} \stackrel{d}{=} X_{n,j-1}, \quad \text{for} \quad j \geq 2,$$

where $\stackrel{d}{=}$ means equality in distribution. Thus,

$$\begin{aligned} X_{n_m} &= X_{2^{m+x}} = X_{2^{m+x},1} + \sum_{j=2}^{\infty} X_{2^{m+x},j} \\ &\stackrel{d}{=} X_{2^{m+x},1} + \sum_{j=1}^{\infty} X_{2^{m-1+x},j} \\ &= X_{2^{m+x},1} + X_{n_{m-1}}, \end{aligned}$$

where

$$\mathsf{Pr}(X_{2^{m+x},1} = 0) = e^{-2^{m-1+x}} \quad \text{and} \quad \mathsf{Pr}(X_{2^{m+x},1} = 1) = 1 - e^{-2^{m-1+x}}.$$

Now let $Y_m = X_{n_m} - m$. Then

$$\begin{aligned} Y_m &= X_{n_m} - m = X_{2^{m+x},1} - 1 + X_{n_{m-1}} - (m-1) \\ &= X_{2^{m+x},1} - 1 + Y_{m-1} = \ldots = \sum_{j=1}^{m} (X_{2^{j+x},1} - 1) + Y_0 \end{aligned}$$

In particular, $Y_{m+1} \leq Y_m$ so that (Y_m) converges in distribution to Y where

$$Y = Y(x) = \sum_{j=1}^{\infty} (X_{2^{j+x},1} - 1) + Y_0 = \sum_{j=1}^{\infty} (X_{2^{j+x},1} - 1 + X_{0,j}).$$

The higher centered moments of D_n can be obtained by analyzing

$$S_1(s) := \exp\{\ln \tilde{G}_n(e^s) - s\mathsf{E}X_n\},$$

where $\tilde{G}_n(z) = (1-z)\tilde{H}_n(z)$ is the generating function of the p.d.f. of X_n. Since

$$\begin{aligned} S_2(s) := \ln \tilde{G}_n(e^s) &= \sum_{j=1}^{\infty} \ln\left(1 + (e^s - 1)(1 - e^{n/2^j})\right) \\ &= \sum_i \frac{(-1)^{i+1}}{i} (e^s - 1)^i \left\{ \sum_{j=1}^{\infty} (1 - e^{-n/2^j})^i \right\}, \end{aligned}$$

letting $V_i := \sum_{j=1}^{\infty} (1-e^{-n/2^j})^i$ we see that

$$\begin{aligned} V_i &= \sum_{j=1}^{\infty} \left\{ \sum_{k=0}^{i} (-1)^k \binom{i}{k} e^{-kn/2^j} \right\} \\ &= \sum_{j=1}^{\infty} \left\{ \sum_{k=0}^{i} (-1)^k \binom{i}{k} e^{-kn/2^j} - \sum_{k=0}^{i} (-1)^k \binom{i}{k} \right\} \\ &= \sum_{j=1}^{\infty} \left\{ \sum_{k=0}^{i} (-1)^{k+1} \binom{i}{k} \left(1 - e^{-kn/2^j}\right) \right\} \end{aligned}$$

These are dyadic sums, which can be asymptotically evaluated with Mellin transform (see [14, Proposition 2]). One obtains

$$V_i \sim \log n - \frac{1}{2} + \frac{\gamma}{L} + \sum_{k=2}^{i} (-1)^{k+1} \binom{i}{k} \log k + \beta_i(\log n). \tag{8}$$

For instance,

$$\beta_1(\log n) = \frac{1}{L} \sum_{k \in \mathbb{Z} \setminus \{0\}} \Gamma\left(\frac{2ik\pi}{L}\right) e^{-2ik\pi \log n}.$$

Hence,

$$S_2(s) = s[\log n - \frac{1}{2} + \frac{\gamma}{L}] + \sum_{i=2}^{\infty} \frac{(-1)^{i+1}(e^s-1)^i}{i} B_i + \sum_{i=1}^{\infty} \frac{(-1)^{i+1}(z-1)^i}{i} \beta_i,$$

where $B_i = \sum_{k=2}^{i} (-1)^{k+1} \binom{i}{k} \log k$. For instance,

$$\begin{aligned} B_1 &= 0, \\ B_2 &= -1, \\ B_3 &= -3 + \log 3, \\ B_4 &= -6 + 4 \log 3 - \log 4, \\ B_5 &= -10 + 10 \log 3 - 5 \log 4 + \log 5. \end{aligned}$$

In order to derive the constant term in the Fourier expansions (in $\log n$), we consider

$$S_3(e^s) = \exp\left[\sum_{i=2}^{\infty} \frac{(-1)^{i+1}(e^s-1)^i}{i} B_i\right]. \tag{9}$$

¿From this equation, we obtain

$$\begin{aligned} \tilde{\sigma}^2 := \mathsf{var}(D_n) &\sim 1 \qquad \text{(as expected), with no periodic contribution} \\ \tilde{\mu}_3 := \mu_3(D_n) &\sim -3 + 2\log 3 = 0.1699250014\ldots, \\ \tilde{\mu}_4 := \mu_4(D_n) &\sim -2(-5 + 6\log 3 - 3\log 4) = 2.980449991\ldots, \\ \tilde{\mu}_5 := \mu_5(D_n) &\sim -45\log 2 + 70\log 3 - 60\log 4 + 24\log 5 = 1.673649353\ldots. \end{aligned}$$

	Theoretical asymptotic value	Observed value
mean	14.62045856...	14.6052...
variance	1	1.0264...
μ_3	0.1699250014...	.1683...
μ_4	2.980449991...	3.1100...

Table 1: Moments.

The neglected terms are made of periodic functions with small amplitude and of $\mathcal{O}(\frac{1}{n})$ contributions.

For $n = 20000$, we have done a simulation (of $m = 4000$ sets). We obtain the results of Table 1. Notice that the asymptotic value of σ^2, $\tilde{\mu}_3$, $\tilde{\mu}_4$ are near those of a Gaussian r.v. Based on our simulation, Figure 1 gives the $\alpha-$ Transform of D_n centered distribution (observed = circle, asymptotic , $S_2(-\alpha)$, = line). Due to the sensitivity of $\tilde{G}_n$ to the mean, we have chosen to normalize by the *observed* mean.

3 Some asymptotic distributions

In this section, we derive asymptotic distributions for $\mathcal{D}_N$, $\mathcal{E}_N$, $\mathcal{M}_N$.

3.1 The largest part

For the maximum part size M_n we have (k will only be used in the neighbourhood of $\log n/L + \mathcal{O}(1)$)

$$\mathsf{Pr}(M_n \leq k) = \left(1 - \frac{1}{2^k}\right)^n \sim e^{-n/2^k}\left(1 - \left(\frac{n}{2^k}\right)^2 \frac{1}{2n}\right).$$

Now, we proceed as in [36]. Set $\eta = j - \log n/L$. Then, with integer j and $\eta = \mathcal{O}(1)$, the distribution is asymptotically given by the extreme-value DF $\varphi_1(x) := e^{-e^{-x}}$:

$$\mathsf{Pr}[M_n - \log n \leq \eta] \sim e^{-e^{-L\eta}}. \tag{10}$$

The mean of this distribution is given by $\frac{\gamma}{L}$ and the variance by $\frac{\pi^2}{6L^2}$. ¿From this and (10) we deduce, as in [36], that $\mathsf{E}M_n \sim \log n + \frac{1}{2} + \frac{\gamma}{L} + \beta(\log n)$ and, as $n = \lfloor (N+1)/2 \rfloor$,

$$\mathsf{E}\mathcal{M}_N \sim \log N - \frac{1}{2} + \frac{\gamma}{L} + \beta(\log N).$$

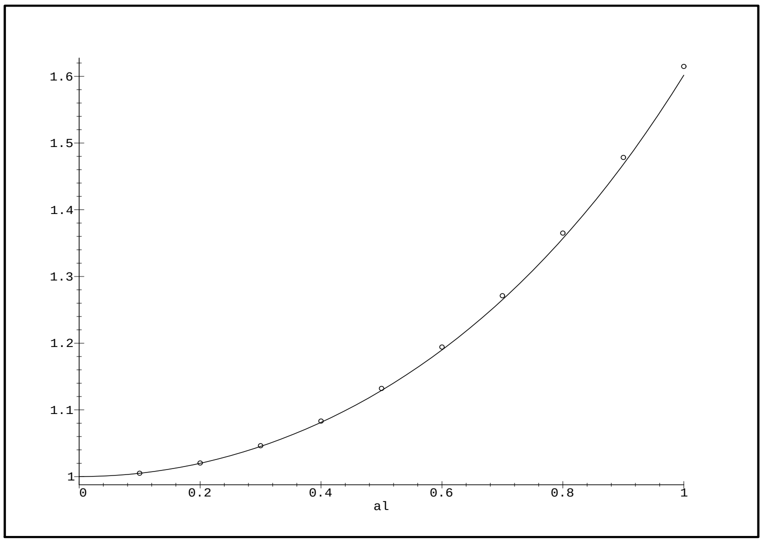

Figure 1: $\alpha-$ Transform of D_n centered distribution

A simulation for $n=$ 20000 of m=4000 sets leads to Figure 2 (observed = circle, asymptotic = line).

The mean of M_n is a special case of a more general result, given by the following Lemma, which relates the moments of a discrete r.v. X to those of the corresponding continuous one. In the sequel, we will consider a discrete r.v. X and a continuous DF $F(x)$ such that $F(x)$ is either an extreme-value DF or a convergent series of such. More general applications conditions are certainly possible, but we will not pursue this matter here. We assume that $\mathsf{Pr}(X - \log n \leq x) \sim F(x)$ in the following sense. Setting $x = j - \log n$ then, with integer j and $x = \mathcal{O}(1)$, $\mathsf{Pr}(X \leq j) \sim F(x), n \to \infty$. Moreover, we assume that the rate of convergence is such that the error is uniformly bounded by a $\mathcal{O}(1/n^\delta)$ term, with $0 < \delta < 1$.

Lemma 3.1 *Let a (discrete) r.v. X be such that $\mathsf{Pr}(X - \log n \leq x) \sim F(x)$, where $F(x)$ is the DF of a continuous r.v. Z with mean m, second moment m_2, variance σ^2 and centered moments (μ_i). Assume that $F(x)$ is either an extreme-value DF or a convergent series of such. Let $\varphi(\alpha) = \mathsf{E}e^{\alpha Z} = e^{\alpha m}\psi(\alpha)$ say, with $\psi(\alpha) = 1 + \frac{\alpha^2}{2}\sigma^2 + \sum_3^\infty \frac{\alpha^i}{i!}\mu_i$. Then the corresponding discrete moments of X are given by*

$$\begin{aligned}
\mathsf{E}(X - \log n) &\sim \int_{-\infty}^{+\infty} x[F(x) - F(x-1)]dx + \beta_2 \\
&= \tilde{m} + \beta_2, \quad \textit{with} \quad \tilde{m} = m + 1/2,
\end{aligned}$$

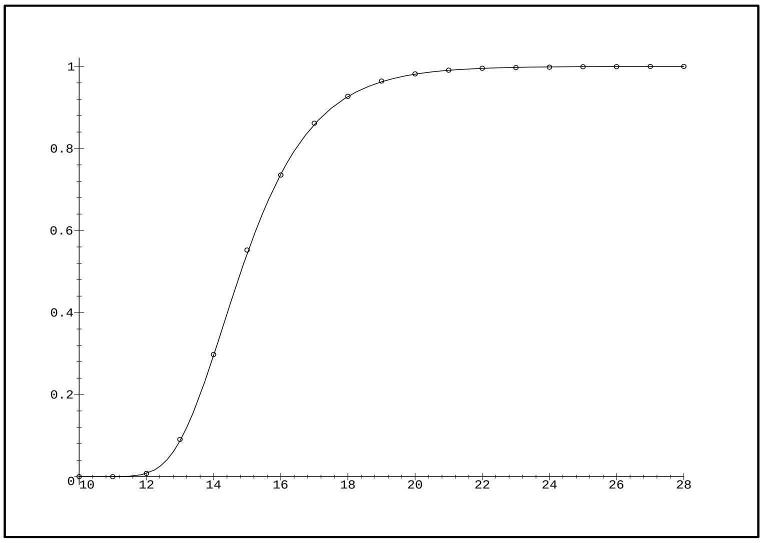

Figure 2: Maximum part size

$$\begin{aligned} \mathsf{var}(X) &\sim \mathsf{E}(X-(\log n+\tilde{m}+\beta_2))^2 \\ &\sim \int_{-\infty}^{+\infty} x^2[F(x)-F(x-1)]dx-\tilde{m}^2+\beta_3 \\ &= m_2+m+1/3-\tilde{m}^2+\beta_3=\tilde{\sigma}^2+\beta_3, \quad \textit{with} \quad \tilde{\sigma}^2=\sigma^2+1/12. \end{aligned}$$

More generally, the centered moments of X are asymptotically given by $\tilde{\mu}_i+\beta_i$, where

$$\Theta(\alpha) := 1+\sum_2^{\infty}\frac{\alpha^i}{i!}\tilde{\mu}_i=\frac{2}{\alpha}sh(\alpha/2)\psi(\alpha).$$

Proof

$$\mathsf{E}(X-\log n)\sim\sum_i[F(i-\log n)-F(i-\log n-1)][i-\log n]. \tag{11}$$

Set $y=2^{-x}$ and $G(y)=F(x)$. Equ.(11) becomes

$$\sum_i[G(n/2^i)-G(n/2^{i+1})][-\log(n/2^i)].$$

This is an harmonic sum and, by Mellin, this leads to (see, for instance, Flajolet [11] for a detailed analysis)

$$\int_0^{\infty}[G(y)-G(y/2)](-\log y)\frac{dy}{Ly}+\beta_2$$

$$\begin{aligned}
&= \int_{-\infty}^{+\infty} [F(x) - F(x-1)]x\ dx + \beta_2 = \int_{-\infty}^{0} F(x)x\ dx - \int_0^{\infty} (1-F(x))x\ dx \\
&- \int_{-\infty}^{0} F(x-1)x\ dx + \int_0^{\infty} (1-F(x-1))x\ dx + \beta_2.
\end{aligned}$$

Set $x-1=y$. We obtain $\int_{-1}^0 (y+1)dy - \int_{-\infty}^0 F(y)dy + \int_0^\infty (1-F(y))dy = m+1/2$. Similarly,

$$\begin{aligned}
\mathsf{var}(X) &\sim \sum_i [G(n/2^i) - G(n/2^{i+1}][-\log(n/2^i) - \tilde{m} - \beta_2]^2 \\
&= \sum_i [G(n/2^i) - G(n/2^{i+1}] \left\{[-\log(n/2^i) - \tilde{m}]^2 - 2[-\log(n/2^i) - \tilde{m}]\beta_2 + \beta_2^2\right\}.
\end{aligned}$$

The first bracket leads to

$$\int_{-\infty}^{+\infty} [F(x) - F(x-1)](x-\tilde{m})^2 dx + \beta_3$$

Now $m_2 = -2\int_{-\infty}^0 F(x)xdx + 2\int_0^\infty (1-F(x))x\ dx$ and

$$\begin{aligned}
\int_{-\infty}^{+\infty} [F(x) - F(x-1)]x^2 dx &= \int_{-\infty}^{0} F(x)x^2 dx - \int_0^{\infty} (1-F(x))x^2 dx \\
&- \int_{-\infty}^{0} F(x-1)x^2 dx + \int_0^{\infty} (1-F(x-1))x^2 dx \\
&= \int_{-1}^{0} (y+1)^2 dy - \int_{-\infty}^{0} F(y)2y\ dy + \int_0^{\infty} (1-F(y))2y dy \\
&- \int_{-\infty}^{0} F(y)dy + \int_0^{\infty} (1-F(y))dy = 1/3 + m_2 + m.
\end{aligned}$$

More generally, let us define $\phi(\alpha) := \int e^{\alpha x}[F(x) - F(x-1)]\ dx$, we obtain

$$\phi(\alpha) = \varphi(\alpha)\frac{e^\alpha - 1}{\alpha} = e^{\alpha m}\frac{e^\alpha - 1}{\alpha}\psi(\alpha).$$

Now $\phi(\alpha) = e^{\alpha\tilde{m}}\Theta(\alpha)$, which proves the lemma. ■

Numerous applications of this lemma can be found in algoritm analysis: let us mention approximate counting ([11], [15]), Tries ([33]), adaptative sampling ([35]), Digital search trees ([34]), leader election ([10]), Lempel-Ziv algorithm ([38]), polyominos analysis ([36]), data structures maxima ([28]), etc. For instance, we derive

$$\begin{aligned}
\tilde{\mu}_3 &= \mu_3, \\
\tilde{\mu}_4 &= \mu_4 + \sigma^2/2 + 1/80, \\
\tilde{\mu}_5 &= \mu_5 + 5/6\mu_3.
\end{aligned}$$

Also

$$\mathsf{var}(M_n) \sim \frac{\pi^2}{6L^2} + \frac{1}{12} + \beta(\log n).$$

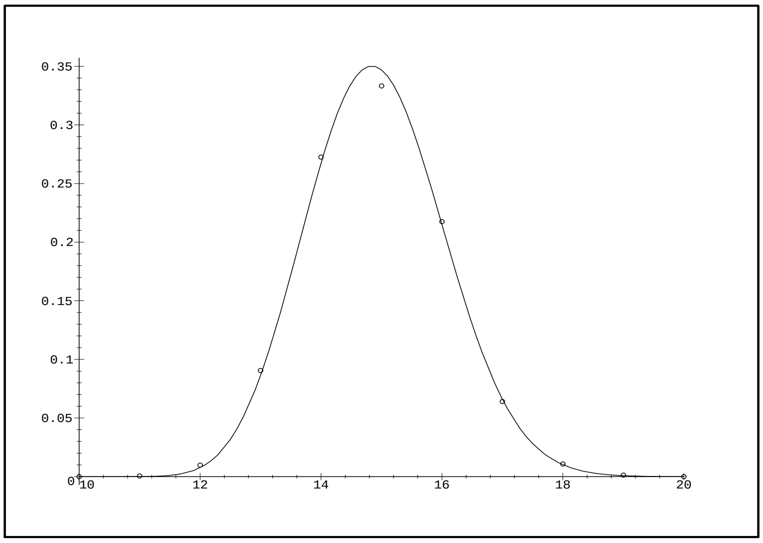

Figure 3: First empty part

3.2 First empty part value

Another variable of interest, $\mathcal{E}_n$, is the first k such that $I_k = 0$, i.e. we are interested in the probability

$$\mathsf{Pr}(\mathcal{E}_n = k) = \mathsf{Pr}(I_i = 1, i = 1 \cdots k-1, I_k = 0).$$

This probability is asymptotically given by

$$\mathsf{Pr}(E_n = k) = \prod_{i=1}^{k-1}(1 - e^{-n/2^i})e^{-n/2^k}.$$

We set $k = \log n + \eta$. This equation leads asymptotically to

$$\varphi(\eta) = e^{-e^{-L\eta}} \prod_{1}^{\infty}[1 - e^{-e^{-L(\eta-i)}}]. \tag{12}$$

Our simulation leads to Figure 3 (observed = circle, asymptotic = line). It is interesting to compare (12) with the corresponding function related to the maximum part size, given from (10) by $e^{-e^{-L\eta}} - e^{-e^{-L(\eta-1)}}$. This is shown in Figure 4, where we see that E_n must have a very concentrated distribution.

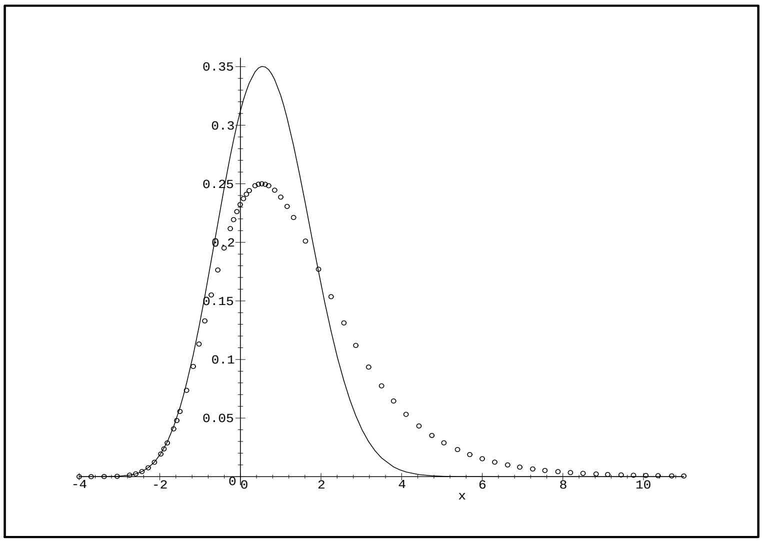

Figure 4: First empty part and maximum part size(circle)

3.3 More on the asymptotic distribution of D_n

If we want to obtain a DF $F(x)$ such that $\mathsf{Pr}[D_n \leq \log n + x] \sim F(x)$, we can use Lemma 3.1, which gives $\tilde{m} = C = m + 1/2$, hence $m = \gamma/L - 1$, and

$$\psi(\alpha) = \frac{\alpha}{2sh(\alpha/2)} S_2(\alpha).$$

For instance

$$\begin{aligned} \sigma^2 &= 11/12, \\ \mu_3 &= \tilde{\mu}_3, \\ \mu_4 &= \tilde{\mu}_4 - 113/240, \\ \mu_5 &= \tilde{\mu}_5 - 5/6\tilde{\mu}_3. \end{aligned}$$

A numerical estimation of $f(x) := F(x) - F(x-1)$ could be obtained by computing the Laplace transform and inverting it numerically, but this is not efficient. It is easier to proceed as follows. We have $\mathsf{Pr}(D_n = j) \sim f(j - \log n)$. Hence, neglecting again the β_j functions, $\mathsf{E}D_n \sim \sum_j f(j - \log n)j$ or $C \sim \sum f(j - \log n)(j - \log n)$. This can be rewritten as

$$C \sim \sum f(j - \lfloor \log n \rfloor - \{\log n\})(j - \lfloor \log n \rfloor - \{\log n\}),$$

or, setting $i := j - \lfloor \log n \rfloor, \theta := \{\log n\}$,

$$C \sim \sum f(i - \theta)(i - \theta), \quad \text{i.e.}$$

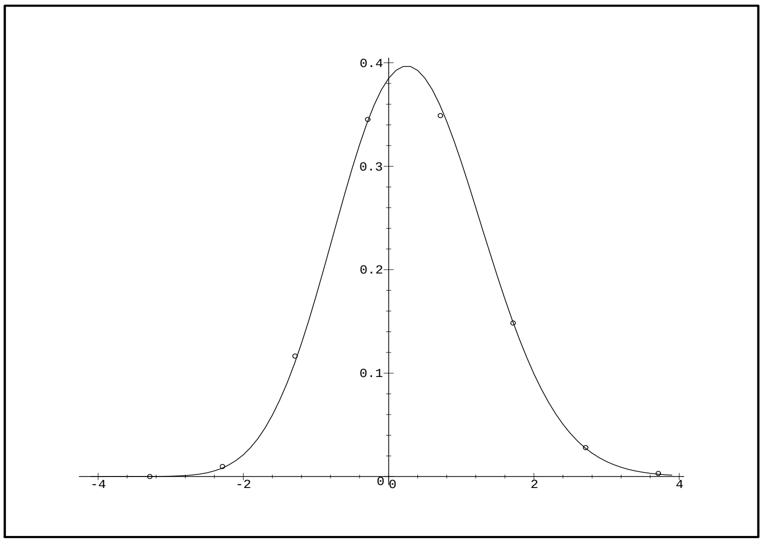

Figure 5: Observed discrete D_n distribution and numerical estimation of f

$$0 \sim \sum f(i-\theta)(i-\theta-C).$$

Similarly

$$\mathsf{var}(D_n) \sim \sum f(i-\theta)(i-\theta-C)^2.$$

Also, $1 \sim \sum f(i-\theta)$. So, we can estimate, for some even d, $[f(-d/2-\theta), \ldots, f(d/2-\theta)]$. We construct a matrix $A[1,\ldots,d+1,1,\ldots,d+1]$ such that

$$A(i,j) := (-d/2-\theta-C+(j-1))^{(i-1)},$$

a column vector b such that $b(1)=1, b(2)=0, b(i)=\tilde{\mu}_{i-1}, i=3,\ldots,d+1$, a column vector x such that $x(i)=f(-d/2-\theta+(i-1))$ after the computation. We solve the systems $Ax=b$ for a set of θ values. For instance, with $d=16, \theta=[0,-1/10,\ldots,-9/10]$, we have obtained a precision of 10^{-4} for f. Figure 5 shows the observed (discrete) probability distribution of D_n together with the numerical estimation of f. The adjustment is quite good. Now we turn to the explicit form of f. The analysis is rather similar to the one we used in [34]. We take the advantage of the fact that all sizes occur *before* the first empty size. That is $D_n \geq E_n - 1$ so that letting $\nu_n = \inf\{k:\ X_{n,k}=0\}$ we have

$$\begin{aligned}\mathsf{Pr}(D_n=m) &= \mathsf{Pr}(D_n=m, E_n \leq m+1) \sim \mathsf{Pr}(X_n=m, \nu_n \leq m+1)\\ &= \sum_{u\geq 0} \mathsf{Pr}(\nu_n = m+1-u, \sum_{r\geq m+2-u} X_{n,r}=u)\end{aligned}$$

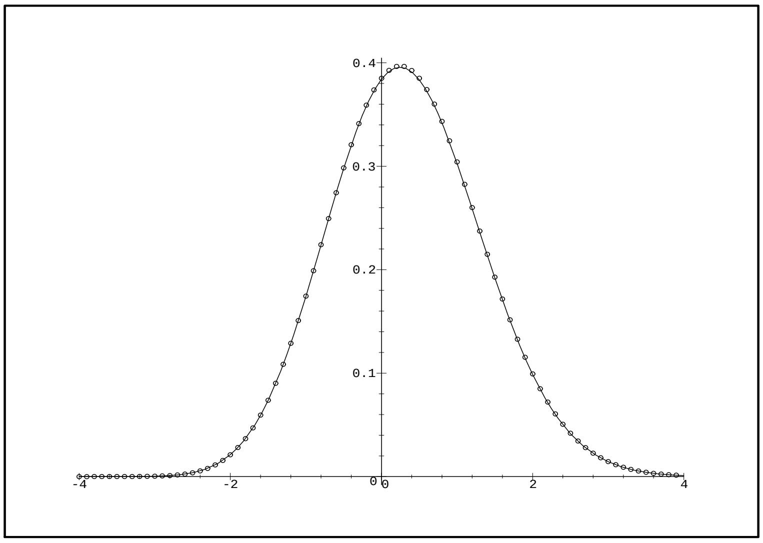

Figure 6: limiting discrete D_n distribution and numerical estimation of f (circle)

$$= \sum_{u \geq 0} \mathsf{Pr}(\nu_n = m + 1 - u) \prod_{r \geq m+2-u} e^{-n/2^r} \sum_{\substack{r_1 \neq \ldots \neq r_u \\ r_j \geq m+2-u}} \frac{1 - e^{-n/2^{r_i}}}{e^{-n/2^{r_i}}}$$

Now set $m = \log n + \eta$ and $r_j = \log n + \eta + w_j$. We obtain the following result:

Theorem 3.2 *With m integer and $\eta = \mathcal{O}(1)$,*

$$\mathit{Pr}(D_n = m) \sim f(\eta) = \sum_{u=0}^{\infty} \varphi(\eta - u + 1) e^{-e^{-L(\eta+1-u)}} \sum_{\substack{w_1 \neq \ldots \neq w_u \\ w_j \geq 2-u}} \prod_{i=1}^{u} \frac{1 - e^{-e^{-L(\eta+w_i)}}}{e^{-e^{-L(\eta+w_i)}}}.$$

$\mathit{Pr}(D_n \leq m) \sim F(\eta)$, *with* $F(\eta) := \sum_0^\infty f(\eta - i)$.

Figure 6 shows the limiting (discrete) probability distribution of D_n, $f(\eta)$, together with the numerical estimation of f (circle).

4 Carlitz Compositions

The Carlitz compositions ([7]) are characterized by the property that two successive parts are different. In this section, we first analyze the hitting probability to a large part value. This allows us to derive the asymptotics for the expected number of distinct part sizes of Carlitz composition, a result first proved in [17]. Then we consider the correlation between two values and obtain the asymptotics of the variance.

4.1 Some known results on Carlitz compositions

In this section, we recall some known asymptotic results on the stochastic description of Carlitz compositions of a large integer N. In [37], we used singularity analysis, based on theorems of Bender, Flajolet, Odlyzko, Soria and Hwang (see [5], [13], [12], [22]). This led to the following results. According to [37, section 2.1], the trivariate generating function of the number of Carlitz compositions of N into m parts with the last part having size i (marked by z, w, and θ, respectively) is given (for a fixed first part size j) by

$$\phi(w,\theta,z|j) = A_2(w,\theta,z|j) + A_1(w,\theta,z)D_2 \tag{13}$$

where

$$\begin{aligned}
A_1(w,\theta,z) &= \sum_{j=1}^{\infty}(-1)^{j+1}\frac{z^j\theta w^j}{1-z^j\theta} \\
D_2 &:= \phi(w,1,z|j) = A_2(w,1,z|j)/h(w,z) \\
A_2(w,\theta,z|j) &= \theta^j w z^j/(1+wz^j) \\
h(w,z) &:= 1+\sum_1^{\infty}\frac{(-1)^j z^j w^j}{1-z^j}.
\end{aligned}$$

Singularity analysis leads to the following results. The number of parts $\mathcal{P}_N$ is asymptotically Gaussian ([37, Theorem 2.1]):

$$\frac{\mathcal{P}_N - N\mu_1}{\sqrt{N}\sigma_1} \stackrel{d}{\sim} \mathcal{N}(0,1), \quad N\to\infty, \tag{14}$$

where

$$\begin{aligned}
\mu_1 &:= -r_1/z^*, \\
\sigma_1^2 &:= \mu_1^2 - r_2/z^*, \\
r_1 &:= -h_w/h_z, \\
r_2 &:= -(r_1^2 h_{zz} + 2r_1 h_{zw} + h_w + h_{ww})/h_z,
\end{aligned}$$

and we denote by z^* the root of $h(1,z)$, i.e $z^* = .57134979315808764311\ldots$ and set $w=1, z=z^*$ in r_1, r_2.
z^* also satisfies

$$\sum \frac{z^{*i}}{1+z^{*i}} = 1. \tag{15}$$

The part sizes are asymptotically given by a Markov Chain (MC):

$$\Pi(i,j) = \frac{z^{*j}(1+z^{*i})}{(1+z^{*j})}, \quad j\neq i \tag{16}$$

Following in details Bender's analysis, we can check that this asymptotic is valid for $i,\ j=\mathcal{O}(\log N)$. Due to its geometrically decreasing tail, the chain is strongly ergodic. It is also reversible. The stationary distribution of (16) is given by

$$\pi(i) = -\frac{z^{*i}}{(1+z^{*i})^2 h_w(1,z^*)}.$$

Set $C_2 := -1/h_w(1,z^*)$=1.3016594836....

4.2 Hitting Probability

In this section, we obtain a precise asymptotic equivalent for the hitting time probability to a large part value. This will be needed in the mean and variance asymptotics. Our results are based on a detailed study of the MC and on some perturbation analysis. Let us analyze the hitting time to some fixed k, $k \gg 1$. This amounts to making k an absorbing state for (16). To study the resulting transient MC $\tilde{\mathbf{\Pi}}$, we assume that we can apply standard perturbation analysis. We show in Appendix A.4 that we can indeed use all parameters in the form given hereafter.

Set $\varepsilon := z^{*k}$. First, we derive $\pi(k) = C_2\varepsilon - 2C_2\varepsilon^2 + \mathcal{O}(\varepsilon^3), \Pi(l,k) = (1+z^{*l})\varepsilon + \mathcal{O}(\varepsilon^2)$. The dominant eigenvalue of $\tilde{\mathbf{\Pi}}$ is given by $1-\gamma_1\varepsilon-\gamma_2\varepsilon^2+\mathcal{O}(\varepsilon^3)$. The corresponding right-eigenvector is given by $\tilde{R}(l) = 1-\alpha(l)\varepsilon+\mathcal{O}(\varepsilon^2)$, the left-eigenvector is given by $\tilde{\pi}(l) = \pi(l)-\delta(l)\varepsilon+\mathcal{O}(\varepsilon^2)$. We shall also compute $\alpha(k)$, notwithstanding the fact that it has no direct probabilistic interpretation. We use the normalization $\boldsymbol{\pi}\tilde{\mathbf{R}} = 1$ (including the $k-$term). We have, with $\sum_l^- := \sum_{l\neq k}$,

$$\sum_m \Pi(l,m)^-[1-\alpha(m)\varepsilon+\mathcal{O}(\varepsilon^2)] = [1-\gamma_1\varepsilon-\gamma_2\varepsilon^2+\mathcal{O}(\varepsilon^3)][1-\alpha(l)\varepsilon+\mathcal{O}(\varepsilon^2)] \quad (17)$$

or

$$1-\varepsilon(1+z^{*l})[l\neq k]-\varepsilon(\mathbf{\Pi}\boldsymbol{\alpha})_l+\mathcal{O}(\varepsilon^2) = 1+\varepsilon[-\gamma_1-\alpha(l)]+\mathcal{O}(\varepsilon^2).$$

The ε term leads to

$$-(1+z^{*l})[l\neq k]-(\mathbf{\Pi}\boldsymbol{\alpha})_l = -\gamma_1-\alpha(l) \quad (18)$$

or

$$[(I-\mathbf{\Pi})\boldsymbol{\alpha}]_l = (1+z^{*l})[l\neq k]-\gamma_1. \quad (19)$$

To derive γ_1, we premultiply (19) by $\boldsymbol{\pi}$, this gives $\sum_l^- \pi(l)(1+z^{*l})-\gamma_1$. Hence $\gamma_1 = C_2$, by (15).

Set $\mathbf{z}^* := \left(z^{*l}\right)$, (column vector). To obtain $\boldsymbol{\alpha}$, we start from (19), which leads to

$$\boldsymbol{\alpha} = \mathbf{M}^-(\mathbf{1}+\mathbf{z}^*)+C_3.\mathbf{1} \quad (20)$$

where $C_3 = \boldsymbol{\pi\alpha} = 0, \mathbf{M} := \sum_{n\geq 0}(\mathbf{\Pi}^n - \mathbf{1}\times\boldsymbol{\pi})$. $\mathbf{M}$ is the Drazin inverse of $\mathbf{I}-\mathbf{\Pi}$. We refer to Campbell and Meyer [6] for a detailed definition and analysis of the Drazin inverse. We have $\mathbf{M} = \mathbf{Z}-\mathbf{1}\times\boldsymbol{\pi}$ where $\mathbf{Z} := [\mathbf{I}-\mathbf{\Pi}+\mathbf{1}\times\boldsymbol{\pi}]^{-1} =$

$\sum_{n\geq 0}[\mathbf{\Pi} - \mathbf{1}\times\boldsymbol{\pi}]^n$ is the potential used in Kemeny, Snell and Knapp [27]. Let us now turn to γ_2. Premultiplying (17) by $\boldsymbol{\pi}$ leads to

$$-[C_2\varepsilon - 2C_2\varepsilon^2 + \mathcal{O}(\varepsilon^3)][1-\alpha(k)\varepsilon + \mathcal{O}(\varepsilon^2)] = [-\gamma_1\varepsilon - m_2\varepsilon^2 + \mathcal{O}(\varepsilon^3)][1-\varepsilon\boldsymbol{\pi\alpha} + \mathcal{O}(\varepsilon^2)]. \tag{21}$$

The ε term leads of course to $\gamma_1 = C_2$. The ε^2 term leads to

$$2C_2 + C_2\alpha(k) = -\gamma_2.$$

Hence

$$\gamma_2 = -2C_2 - C_2\alpha(k) = -2C_2 - C_2(\mathbf{M}^-(\mathbf{1}+\mathbf{z}^*))_k, \quad \text{by (20)}.$$

But, from (43), $(\mathbf{M}^-(\mathbf{1}+\mathbf{z}^*))_k = C_4 + \mathcal{O}(\varepsilon)$, with $C_4 = -1.2774603654\ldots$ so $\gamma_2 = -C_2(2+C_4)$. With (38), we also derive

$$\begin{aligned} \alpha(k) &= C_4 + \mathcal{O}(\varepsilon), \\ \alpha(l) &= C_4 + 1 + \mathcal{O}(z^{*l}), \quad l \gg 1, \quad l \neq k. \end{aligned} \tag{22}$$

$\boldsymbol{\delta}$ could be similarly computed, but we will not need its explicit form. However, we will use the normalization $\tilde{\boldsymbol{\pi}}^-\tilde{\mathbf{R}} = 1$, which gives

$$-C_2 - \boldsymbol{\delta}\mathbf{1} = 0. \tag{23}$$

By Keilson [26], Aldous [1], Aldous and Brown, [2], [3], we know that the hitting time to a distant state is asymptotically exponential. But, here, we need a precise equivalent. For further use, set $C_9 := -\gamma_2 - \gamma_1^2/2$. Let us first fix n. We analyze, with $k = \Theta(\log n)$ and starting with the stationary distribution,

$$\begin{aligned} \mathsf{Pr}(T_k > n) &= \mathsf{Pr}_\pi^{(n)}(I_k = 0) = \boldsymbol{\pi}^-(\mathbf{\Pi}^-)^{n-}\mathbf{1} \\ &\sim \boldsymbol{\pi}^-[\mathbf{1} - \varepsilon\boldsymbol{\alpha} + \mathcal{O}(\varepsilon^2)][\boldsymbol{\pi} - \boldsymbol{\delta}\varepsilon]^-\mathbf{1}[1 - \gamma_1\varepsilon - \gamma_2\varepsilon^2 + \mathcal{O}(\varepsilon^3)]^n \\ &\sim \boldsymbol{\pi}^-[\mathbf{1} - \varepsilon\boldsymbol{\alpha} + \mathcal{O}(\varepsilon^2)][1 + \varepsilon(-C_2 - \boldsymbol{\delta}\mathbf{1})]e^{-n\gamma_1\varepsilon}[1 + n\varepsilon^2 C_9 + \mathcal{O}(n\varepsilon^3)] \\ &\sim \boldsymbol{\pi}^-[\mathbf{1} + \varepsilon(-C_2\mathbf{1} - (\delta\mathbf{1})\mathbf{1} - \boldsymbol{\alpha}) + \mathcal{O}(\varepsilon^2)]e^{-n\gamma_1\varepsilon}[1 + n\varepsilon^2 C_9 + \mathcal{O}(n\varepsilon^3)] \\ &\sim [1 - \varepsilon C_2 + \mathcal{O}(\varepsilon^2)]e^{-n\gamma_1\varepsilon}[1 + n\varepsilon^2 C_9 + \mathcal{O}(n\varepsilon^3)], \quad \text{by (23)}. \end{aligned}$$

Now, we set $\tilde{n} = nC_2$, which leads to the following result

Lemma 4.1

$$\mathit{Pr}_\pi^{(n)}(I_k = 0) \sim \left[1 - \frac{(\tilde{n}\varepsilon)}{\tilde{n}}C_2 + \frac{(\tilde{n}\varepsilon)^2}{\tilde{n}}C_{10} + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\tilde{n}\varepsilon^3)\right] e^{-\tilde{n}\varepsilon} \tag{24}$$

with $C_{10} := C_9/C_2$.

We finally obtain $\bar{V}_1 \sim \ln\tilde{n}/\tilde{L} + \tilde{C}$, with $\tilde{C} = -1/2 + \gamma/\tilde{L}$ and $\tilde{L} = -\ln(z^*)$. But actually, n is a r.v. representing the number of parts. So, according to (14), we derive

$$\mathsf{E}(\mathcal{D}_N) \sim \int e^{-(n-N\mu_1)^2/(N\sigma_1^2)}\bar{V}_1(n)\, dn$$

so, asymptotically, we must replace $\ln(n)$ by $\ln(N)+\ln(h_w/(h_z z^*))$, hence we replace $\ln(\tilde{n})$ by $\ln(N)-\ln(-h_z)-\ln(z^*)$ and finally,

$$\mathsf{E}(\mathcal{D}_N) \sim \ln(N)/\tilde{L} - \ln(-h_z)/\tilde{L} + 1/2 + \gamma/\tilde{L} + \beta(\log N) + \mathcal{O}(1/N).$$

So we recover, by another method, a result first proved in [17]. Similarly, in (8), we use $\ln k/\tilde{L}$ instead of $\log_2 k$. For instance,

$$\begin{aligned} B_2 &= -\ln 2/\tilde{L}, \\ B_3 &= -3\ln 2/\tilde{L} + \ln 3/\tilde{L}, \\ B_4 &= -6\ln 2/\tilde{L} + 4\ln 3/\tilde{L} - \ln 4/\tilde{L}, \\ B_5 &= -10\ln 2/\tilde{L} + 10\ln 3/\tilde{L} - 5\ln 4/\tilde{L} + \ln 5/\tilde{L}. \end{aligned}$$

4.3 Correlations and variance

In this section, we first consider the correlation between I_k and I_j, and we finally obtain the asymptotics of $\mathsf{var}(\mathcal{D}_N)$. Fix $k<j$, $k \gg 1$. Set $\tau = z^{*(j-k)}$. $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, γ_1, γ_2 depend now on k and j and $\sum_l^- := \sum_{l \neq k,j}$. Equ.(19) becomes

$$[(\mathbf{I}-\mathbf{\Pi})\boldsymbol{\alpha}]_l = (1+z^{*l})[l \neq k,j](1+\tau) - \gamma_1.$$

Hence, $\gamma_1 = C_2(1+\tau)$ and $\boldsymbol{\alpha} = (1+\tau)\mathbf{M}^-(\mathbf{1}+\mathbf{z}^*) + C_{11}$, with $C_{11} = \boldsymbol{\pi\alpha} = 0$. Instead of (23), we have

$$-C_2(1+\tau) - \boldsymbol{\delta}\mathbf{1} = 0.$$

Equ.(21) becomes

$$\begin{aligned} -[C_2\varepsilon(1+\tau) &- 2C_2\varepsilon^2 - 2C_2\varepsilon^2\tau^2 - C_2\varepsilon^2\alpha(k) - C_2\varepsilon^2\tau\alpha(j) + \mathcal{O}(\varepsilon^3)] \\ &= [-\gamma_1\varepsilon - \gamma_2\varepsilon^2 + \mathcal{O}(\varepsilon^3)][1 - \varepsilon\boldsymbol{\pi\alpha} + \mathcal{O}(\varepsilon^2)]. \end{aligned}$$

The ε term leads to $\gamma_1 = C_2(1+\tau)$, as it should. The ε^2 term leads to

$$2C_2(1+\tau^2) + C_2\alpha(k) + C_2\tau\alpha(j) = -\gamma_2.$$

Hence

$$\begin{aligned} \gamma_2 &= -2C_2(1+\tau^2) - C_2(1+\tau)[(M^-(\mathbf{1}+\mathbf{z}^*))_k + \tau(M^-(\mathbf{1}+\mathbf{z}^*))_j] \\ &\sim -2C_2(1+\tau^2) - C_2(1+\tau)[(1+\tau)C_4 - M(k,j) - \tau M(j,k)]. \end{aligned}$$

After some algebra, we obtain

$$\begin{aligned} \mathsf{Pr}_\pi^{(n)}(I_k{=}0 \cap I_j = 0) \sim \Big\{ 1 - (z^{*k}+z^{*j})C_2\frac{\tilde{n}}{\tilde{n}} - \frac{\tilde{n}^2}{\tilde{n}}\frac{1}{C_2}\Big[-2C_2(z^{*2k}+z^{*2j}) \\ -C_2(z^{*k}+z^{*j})^2C_4 + C_2(z^{*k}+z^{*j})[z^{*k}M(k,j) + z^{*j}M(j,k)] \\ +C_2^2/2(z^{*k}+z^{*j})^2\Big]\Big\} e^{-\tilde{n}(z^{*k}+z^{*j})}. \end{aligned}$$

Now we analyze $\mathsf{Pr}[I_i = 1 \cap I_j = 1]$. Consider the case $i = \Theta(\log n)$, $j = \Theta(\log n)$ (other cases are unimportant by the "sum splitting technique" as described in Knuth [31, p.131]).

$$\begin{aligned}
\mathsf{E}(I_i I_j) &= \mathsf{Pr}[I_i = 1 \cap I_j = 1] = 1 - [\mathsf{Pr}[I_i = 0] + \mathsf{Pr}[I_j = 0] - \mathsf{Pr}[I_i = 0 \cap I_j = 0]] \\
\sim\ & (1 - e^{-\tilde{n}z^{*i}})(1 - e^{-\tilde{n}z^{*j}}) \\
+\ & \frac{1}{\tilde{n}} \Bigg\{ - e^{-\tilde{n}z^{*i}} \left[-(\tilde{n}z^{*i})C_2 + (\tilde{n}z^{*i})^2 C_{10} + \mathcal{O}(\tilde{n}z^{*i2}) + \mathcal{O}(\tilde{n}^2 z^{*i3}) \right] \\
-\ & e^{-\tilde{n}z^{*j}} \left[-(\tilde{n}z^{*j})C_2 + (\tilde{n}z^{*j})^2 C_{10} + \mathcal{O}(\tilde{n}z^{*j2}) + \mathcal{O}(\tilde{n}^2 z^{*j3}) \right] \\
+\ & e^{-\tilde{n}(z^{*k}+z^{*j})} \Bigg[- (z^{*k} + z^{*j})C_2\tilde{n} - \frac{\tilde{n}^2}{C_2} \left[-2C_2(z^{*2k} + z^{*2j}) - C_2(z^{*k} + z^{*j})^2 C_4 \right. \\
+\ & \left. C_2(z^{*k} + z^{*j})[z^{*k}M(k,j) + z^{*j}M(j,k)] + C_2^2/2(z^{*k} + z^{*j})^2 \right] \Bigg] \Bigg\}.
\end{aligned}$$

We would like to conclude that all moments of D_n can be based on this lemma and that for instance, $\mathsf{var}(D_n) \sim \bar{V}_1^2 - \bar{V}_2 + \bar{V}_1 - \bar{V}_1^2 + \tilde{\beta} = \ln(2)/\tilde{L} + \tilde{\beta}$, for some periodic function $\tilde{\beta}(\log n)$. But we must carefully check the effect of β's and $\frac{1}{n}$ contributions. Actually, we have the following theorem

Theorem 4.2 *$\mathsf{var}(\mathcal{D}_N) \sim \ln(2)/\tilde{L} + \beta(\log N) + \mathcal{O}(\frac{1}{N})$ for some periodic function $\beta(\log N)$.*

Proof
It is easy to check that

$$\sum_1^\infty e^{-nz^{*i}} nz^{*i} \sim \frac{1}{\tilde{L}} + \beta_4, \qquad \sum_1^\infty e^{-nz^{*i}} (nz^{*i})^2 \sim \frac{1}{\tilde{L}} + \beta_5,$$

and

$$\sum_1^\infty e^{-nz^{*i}} (1 - e^{-nz^{*i}}) \left(nz^{*i}\right)^2 \sim \frac{3}{4\tilde{L}} + \beta_6.$$

Indeed, these are harmonic sums, which again are computed with Mellin Transforms.
$S_2 := \mathsf{E}(D_n^2) = \mathsf{E}((\sum I_i)^2) = \sum_{i \neq j} \mathsf{E}(I_i I_j) + \sum_i \mathsf{E}(I_i)$. Now, after some tedious algebra,

$$S_1 := \sum_{i \neq j} \mathsf{E}(I_i I_j) \sim (\bar{V}_1 + \beta_1)^2 + \frac{C_{12}}{n\tilde{L}} \bar{V}_1 + \mathcal{O}(1/n) - (\bar{V}_2 + \beta_2) + \beta_7,$$

for some $\beta_7(\log n)$ and some constant C_{12}. Actually, we have to compute the two extra terms

$$\sum_k \sum_j e^{-\tilde{n}(z^{*k}+z^{*j})} \tilde{n}^2 z^{*2j} M(j,k) \qquad \text{and} \tag{25}$$

$$\sum_k \sum_j e^{-\tilde{n}(z^{*k}+z^{*j})} \tilde{n}^2 z^{*k} z^{*j} M(j,k) \tag{26}$$

which are asymptotically constant by (46). Finally,

$$\mathsf{var}(D_n) \sim S_1 + (\bar{V}_1 + \beta_1 + \frac{C_{12}}{2n\tilde{L}}) - (\bar{V}_1 + \beta_1 + \frac{C_{12}}{2n\tilde{L}})^2 + \beta_7 = \ln(2)/\tilde{L} + \tilde{\beta} + \mathcal{O}(\frac{1}{n}).$$

Again, proceeding as in the previous mean analysis, the transfer to $\mathcal{D}_N$ is immediate. ■

4.4 Some perspective

The generating function is given, **under the independence assumption**, by the same expression as for D_n (with L and n replaced by $\tilde{L}$ and $\tilde{n}$). The part size maximum has already been analyzed in [37].

We have also done a series of simulations with the Carlitz MC. The results are quite similar to the GEOM case.

To derive the independence assumption, we could try a model based on Markov chains on urns. Indeed, an alternative proof of (6), can be obtained by using an urn model, as in Sevastyanov and Chistyakov, [39] and Chistyakov, [8], the Poissonization method and the standard saddle-point method (see, for instance, Flajolet and Sedgewick, [16]). This will be the object of future work.

5 Conclusion

Using various techniques from analysis and probability theory, we have analyzed the stochastic properties of the distinctness of classical compositions. The mean and variance have been derived for the Carlitz case. An open problem is to prove the independence assumption in the latter case, which is corroborated by our simulations.

Acknowledgements
We thank H. Hwang for useful comments on several preliminary versions of this paper. The pertinent comments of the referees led to substantial improvements in the presentation.

References

[1] Aldous, D. *Probability Approximations via the Poisson Clumping Heuristics*, Springer–Verlag, Heidelberg, 1989.

[2] Aldous, D., Brown, M. Inequalities for rare events in time–reversible Markov chains I. *Stochastic Inequalities*, IMS **22**(1992), 1–16.

[3] Aldous, D., Brown, M. Inequalities for rare events in time–reversible Markov chains II. *Stochastic Process. Appl.* **44**(1993), 15–25.

[4] Andrews, G. E. *The Theory of Partitions*, Addison – Wesley, Reading, MA, 1976.

[5] Bender, E.A. Central and local limit theorems applied to asymptotic enumeration, *J. Combin. Theor. Ser. A*, **15**, 91-111 (1973).

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

[7] Carlitz, L. Restricted Compositions, *The Fibonacci Quart.*, **14**, 254-264, 1976.

[8] Chistyakov, V. P. Discrete limit distributions in the problem of balls falling in cells with arbitrary probabilities, *Math. Notes.* **1**(1967), 6–11.

[9] Erdös, P., Lehner, J. The distribution of the number of summands in the partitions of positive integer, *Duke Math. J.* **8** (1941), 335–345.

[10] Fill, J. A., Mahmoud, H., Szpankowski, W. On the distribution of the duration of a randomized leader election algorithm, *Ann. Appl. Probab.* **1**(1996), 1260 – 1283.

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

[12] Flajolet, P., Odlyzko, A. Singularity analysis of generating functions, *SIAM J. Algebraic and Discrete Methods* , **3**, 216-240 (1990).

[13] Flajolet, P., Soria, M. General combinatorial schemes: Gaussian limit distribution and exponential tails, *Discrete Math.*, **114**, 159-180 (1993).

[14] Flajolet, P., Gourdon, X., Dumas, P. Mellin transform and asymptotics: Harmonic sums, *Theoret. Comput. Sci.*, **144** (1995), 3–58.

[15] Flajolet, P., Martin, G. Probabilistic counting algorithms for data base applications, *J. Comput. System Sci.* **31**(1985), 182–209.

[16] Flajolet, P., Sedgewick, R. The Average Case Analysis of Algorithms: Saddle Point Asymptotics , *INRIA T.R.* **2376**(1994).

[17] Goh, W. M. Y., Hitczenko, P. On the number of distinct part sizes in a random Carlitz composition, *Europ. J. Combin.*, to appear (http://www.mcs.drexel.edu/∼phitczen).

[18] Goh, W. M. Y., Schmutz, E. The number of different part sizes in a random integer partition, *J. Combin. Theory Ser. A* **69** (1995), 149–158.

[19] Hitczenko, P., Savage, C. D. On the multiplicity of parts in a random composition of a large integer, preprint, 1999 (http://www.mcs.drexel.edu/∼phitczen).

[20] Hitczenko, P., Stengle, G. Expected number of distinct part sizes in a random integer composition, *Combin. Probab. Comput.* **9**(2000), 519–527, (also available at http://www.mcs.drexel.edu/∼phitczen).

[21] Hwang, H. K., Yeh, Y. N. Measures of distinctness for random partitions and compositions of an integer, *Adv. Appl. Math.* **19**(1997), 378 – 414.

[22] Hwang, H.K. On convergence rates in the central limit theorems for combinatorial structures, *European J. Combin.*, **19**, 329-343 (1998).

[23] Jacquet, S., Szpankowski, W. Analytical de–Poissonization and its applications, *Theoret. Comput. Sci.* **201** (1998), 1 – 62.

[24] Janson, S., Szpankowski, W. Analysis of the asymmetric leader election algorithm, *Electron. J. Combin.* **4** (1997), no. 1, research paper 17, 16pp.

[25] Kato, T. *Perturbation Theory for Linear Operators*, Springer–Verlag, Heidelberg, 1966.

[26] Keilson, J. *Markov Chain Models – Rarity and Exponentiality*, Springer–Verlag, Heidelberg, 1979.

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

[28] Kenyon, C., Louchard, G., Schott, R. Data structures maxima. *SIAM J. Comput.* **26** (1997), 1006–1042.

[29] Knopfmacher, A., Mays, H. Compositions with m distinct parts, *Ars Combin.* **53** (1999), 111–128.

[30] Knopfmacher, A., Prodinger, H. On Carlitz compositions *European J. Combin.* **19** (1998), 579–589 (http://www.idealibrary.com).

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

[32] Louchard, G. Recurrence times and capacities for finite ergodic chains. *Duke Math. J.* **33**(1966), 13–22.

[33] Louchard, G. Brownian motion and algoritm complexity, *BIT* **26**(1986), 17–34.

[34] Louchard, G. Exact and asymptotic distributions in digital and binary search trees. *RAIRO Inform. Théor. Appl.* **21**(1987), 479–495.

[35] Louchard, G. Probabilistic analysis of adaptive sampling, *Random Structures Algorithms* **10**(1997), 157–168.

[36] Louchard, G. Probabilistic analysis of column–convex and directed diagonally–convex animals II, *Random Structures Algorithms* **15**(1999), 1–23.

[37] Louchard, G., Prodinger, H. Probabilistic analysis of Carlitz compositions, preprint (http://www.wits.ac.za/helmut/paperlst.htm).

[38] Louchard, G., Szpankowski, W. Average profile and limiting distribution for a phrase size in the Lempel–Ziv parsing algorithm, *IEEE Trans. Inform. Theory* **41**(1995), 478–488.

[39] Sevastyanov, B.A., Chistyakov, V.P. Asymptotic normality in the classical problem of balls, *Theor. Probab. Appl.* **9** (1964), 198–211.

[40] Wilf, H. S. Three problems in combinatorial asymptotics, *J. Combin. Theory Ser. A* **35** (1983), 199–207.

A Appendix: Some Probabilistic Potential Results

The matrix $\mathbf{M}$ has a lot of structure in it. This is closely related to MC Potential theory (see, for instance, Kemeny, Snell and Knapp, [27] and Louchard [32]). In Appendix A.1-A.3, we provide connections with hitting times, and several first-order approximations of M and of related summations. In A.4, we analyze the perturbation problem. A.5 is devoted to a more analytic view of $\mathbf{M}$, deduced from the trivariate generating function.

A.1 Hitting time

Again we assume, in A1-A3 that we can apply standard perturbation analysis. The justification will be given in A.4. Let us first analyze $h(i) := \mathsf{E}_i[T_k],\ k \gg 1$, where T_k is the hitting time to state k. We have (dropping k to ease the notation)

$$\mathbf{h} = \mathbf{1} + \mathbf{\Pi}^{-}\mathbf{h} = \mathbf{1} + \mathbf{\Pi}\mathbf{h} - \mathbf{\Pi}^{+}\mathbf{h}. \tag{27}$$

Hence

$$\mathbf{h} = -\mathbf{M}\mathbf{\Pi}^{+}\mathbf{h} + C_5(k).\mathbf{1} \tag{28}$$

with $C_5(k) = \boldsymbol{\pi}\mathbf{h}$. But from [32] (the results are obtained for finite MC, but they are easily converted to the denumerable strong ergodic case), we know that

$$\begin{aligned}
\boldsymbol{\pi}^{+}\mathbf{h} &= 1 \quad \text{(This is equivalent to a theorem of Kac)}, && (29)\\
\mathbf{\Pi}\mathbf{M} &= \mathbf{M}\mathbf{\Pi} = \mathbf{M} - \mathbf{I} + \mathbf{1} \times \boldsymbol{\pi}, && (30)\\
\boldsymbol{\pi}\mathbf{M} &= \mathbf{M}\mathbf{1} = 0, && (31)\\
C_6(k) - \mathbf{M}^{+}\mathbf{h} &= 0 \quad \text{on} \quad k, \quad \text{for some constant} \quad C_6(k), && (32)\\
C_6(k) - \mathbf{M}^{+}\mathbf{h} &= h \quad \text{on} \quad j \neq k, &&\\
C_6(k) &= \boldsymbol{\pi}^{-}\mathbf{h}. && (33)
\end{aligned}$$

¿From (29) and (33), we see that $C_6(k) = C_5(k) - 1$ and from (29), with $\varepsilon := z^{*k}$,

$$h(k) = \frac{1}{\pi(k)} = \frac{1}{\varepsilon C_2} + \frac{2}{C_2} + \mathcal{O}(\varepsilon). \tag{34}$$

On the other side, (28) leads to

$$h(k) = C_5(k) - C_4(k)/C_2 + \mathcal{O}(\varepsilon)$$

with

$$C_4(k) := [\mathbf{M}^{-}(\mathbf{1} + \mathbf{z}^{*})]_k. \tag{35}$$

By perturbation analysis, we know that, for ε sufficiently small:

$$h(i) = C_7/\varepsilon + \varphi(i) + \mathcal{O}(\varepsilon).$$

We should write $C_7(i)$, but we will soon check that C_7 is independent of i. Equ.(27) leads to

$$(\mathbf{I}-\mathbf{\Pi})C_7 = 0$$

which confirms that C_7 is independent of i and $C_7 = 1/C_2$ by (34). The independent term leads to

$$[(\mathbf{I}-\mathbf{\Pi})\boldsymbol{\varphi}]_l = 1-(1+z^{*l})[l\neq k]C_7$$

i.e.

$$\varphi(i) = -C_7[\mathbf{M}^-(\mathbf{1}+\mathbf{z}^*)]_i + C_8(k) \tag{36}$$

with

$$C_8(k) = \boldsymbol{\pi\varphi} = -C_7/\varepsilon + C_6(k) + 1 + \mathcal{O}(\varepsilon).$$

For $i = k$, (36) with (34) gives

$$C_6(k) = C_7/\varepsilon + 2C_7 - 1 + C_7C_4(k) + \mathcal{O}(\varepsilon). \tag{37}$$

Then (32) leads to

$$M(k,k) = 1 + M^{(1)}(k,k)\varepsilon + \mathcal{O}(\varepsilon^2) \tag{38}$$

and $C_6(k) = C_7/\varepsilon + 2C_7 + C_7M^{(1)}(k,k) + \mathcal{O}(\varepsilon)$. With (37), we obtain

$$C_4(k) = M^{(1)}(k,k) + C_2 + \mathcal{O}(\varepsilon)$$

which we can also derive from (30).

A.2 Analysis of $C_4(k)$

Let us start from (30). The dominant term of $\mathbf{\Pi M}(k,k)$ is clearly in ε and leads to

$$\sum_{l\neq k}\frac{z^{*l}}{1+z^{*l}}M(l,k) = \varepsilon C_4(k) + \mathcal{O}(\varepsilon^2). \tag{39}$$

With $j\neq k$, we derive

$$\sum_{l\neq j}\Pi(j,l)M(l,k) = M(j,k) + \pi(k)$$

or,

$$\begin{aligned}(1+z^{*j})&\left[\frac{\varepsilon}{1+\varepsilon}(1+\mathcal{O}(\varepsilon)) + \sum_{l\neq k}\frac{z^{*l}}{1+z^{*l}}M(l,k) - \frac{z^{*j}}{1+z^{*j}}M(j,k)\right]\\ &= M(j,k)+\pi(k)\end{aligned} \tag{40}$$

which shows that, with $j = \mathcal{O}(k)$, $\quad j\neq k$,

$$M(j,k) = \varepsilon[C_4(k) - C_2 + 1] + \mathcal{O}(\varepsilon^2) * [j>k] + \mathcal{O}(\varepsilon z^{*j}) * [j<k]. \tag{41}$$

The dominant term is *independent of* j. Note that (31) gives other relations on $\mathbf{M}$: we obtain

$$\begin{aligned}\sum_{l\neq k}\frac{z^{*l}}{(1+z^{*l})^2}M(l,k) &= -\varepsilon+\mathcal{O}(\varepsilon^2)\\ \sum_{l\neq k}M(k,l) &= -1+\mathcal{O}(\varepsilon).\end{aligned}$$

The second form of (30) gives, for $\mathbf{M\Pi}(l,k)$ the relation

$$\sum_{l\neq k}M(k,l)\Pi(l,k)=\varepsilon C_4(k)+\mathcal{O}(\varepsilon^2),\quad \text{as expected.}$$

With $j\neq k$, we obtain

$$\frac{z^{*k}}{1+z^{*k}}\left[\sum_{l\neq j}M(j,l)(1+z^{*l})+M(j,j)(1+z^{*j})-M(j,k)(1+z^{*k})\right] = M(j,k)+\pi(k) \tag{42}$$

and, with (41), $C_4(j)=C_4(k)+\mathcal{O}(\varepsilon)+\mathcal{O}(z^{*j})$, i.e.

$$C_4(k)=C_4+\mathcal{O}(\varepsilon)\text{ for some constant }C_4. \tag{43}$$

A numerical investigation gives $C_4=-1.2774603654\ldots$, and (41) becomes

$$M(j,k)=-1.5791198491\ldots\varepsilon+\mathcal{O}(\varepsilon^2),\quad j\neq k, j=\mathcal{O}(k).$$

Finally, (35) can also be rewritten, with (38), as $C_4=-1+\lim_{k\to\infty}(M\mathbf{z}^*)_k$.

A.3 Analysis of two summations (25) and (26)

Let us first remark that (40) allows a first-order estimation of $M(j,k), k\gg 1, j=\mathcal{O}(1)$. Indeed, this leads to

$$(1+z^{*j})\left[\varepsilon+\varepsilon C_4-\frac{z^{*j}}{1+z^{*j}}M(j,k)\right]=M(j,k)+\varepsilon C_2+\mathcal{O}(\varepsilon^2).$$

Hence

$$M(j,k)=\varepsilon\left[-\frac{C_2}{1+z^{*j}}+1+C_4\right]+\mathcal{O}(\varepsilon^2). \tag{44}$$

The coefficient of ε in the dominant term is *independent of* k. Also by (15), this is compatible with (39) and $-C_2+(1+z^{*j})(1+C_4)<0$.
Similarly, (42) leads, for $j\gg 1, k=\mathcal{O}(1)$, to

$$\frac{z^{*k}}{1+z^{*k}}[C_4+1-M(j,k)(1+z^{*k})]=M(j,k)+\pi(k)+\mathcal{O}(z^{*j}).$$

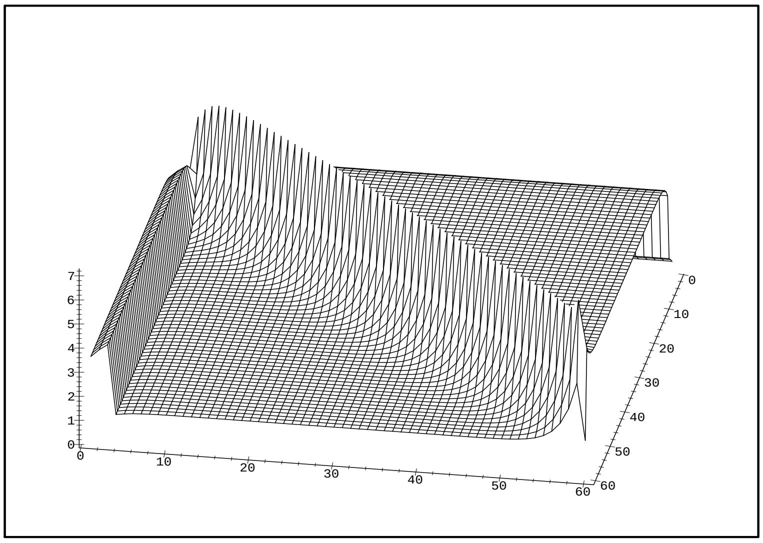

Figure 7: M, after substracting approximations

Hence

$$M(j,k) = \frac{z^{*k}}{(1+z^{*k})^2}\left[-\frac{C_2}{1+z^{*k}} + 1 + C_4\right] + \mathcal{O}(z^{*k}z^{*j}). \tag{45}$$

The dominant term is *independent of* j. Again, by (15) , this is compatible with (35), and $-C_2 + (1+z^{*k})(1+C_4) < 0$.

We have checked the quality of our approximations by substracting, from $\mathbf{M}$, the various expressions given by (38), (41), (44), (45) and normalizing by the suitable ε power. This leads to Figure 7 .

Now we divide the summation in (25) and (26) into 4 regions. Set $d := -\alpha\ln(\tilde{n})/\ln(z^*)$, for some $0 < \alpha < 1$. It is easy to see that the 3 regions: $(j,k) \in [1,\dots,d]\times[1,\dots,d], [1,\dots,d]\times[d+1,\dots,\infty], [d+1,\dots,\infty]\times[1,\dots,d]$ lead to (exponentially) small contribution $\mathcal{O}(e^{-n^{1-\alpha}})$. Moreover, in the last region, $[d+1,\dots,\infty]\times[d+1,\dots,\infty]$, only $M(i,i)$ leads to a $\mathcal{O}(1)$ contribution, given by

$$\sum_{i=d+1}^{\infty} e^{-\tilde{n}2z^{*i}}\tilde{n}^2 z^{*2i} + \mathcal{O}(z^{*d})$$

which can be replaced by

$$\sum_{i=1}^{\infty} e^{-\tilde{n}2z^{*i}}\tilde{n}^2 z^{*2i} + \mathcal{O}(\tilde{n}^{-\alpha}).$$

The sum is again an harmonic sum, with value

$$\frac{1}{4\tilde{L}} + \beta_8. \tag{46}$$

A.4 Perturbation Analysis

First of all let us truncate the MC as follows: we collapse all probability measure from $[\kappa, \ldots, \infty]$ to κ, where $\kappa \gg 1$ and $\kappa \gg k$. We denote by $\mathbf{\Pi}_1$ this new MC and by $\mathbf{\Pi}^*$ the original matrix $\mathbf{\Pi}$ restricted to $\kappa \times \kappa$. This will be a useful tool in the sequel. Set $\eta := z^{*\kappa}$. We see that

$$\begin{aligned}
(\mathbf{\Pi}^* - \mathbf{\Pi}_1)(l, j) &= 0, \quad j \neq \kappa, l \leq \kappa \\
(\mathbf{\Pi}^* - \mathbf{\Pi}_1)(l, \kappa) &= C_{12}(1 + z^{*l})\eta(1 + \mathcal{O}(\eta)) \quad l < \kappa, \\
(\mathbf{\Pi}^* - \mathbf{\Pi}_1)(\kappa, \kappa) &= C_{13}\eta(1 + \mathcal{O}(\eta)).
\end{aligned}$$

Now we formulate 3 remarks:
1. As $\mathbf{\Pi}_1$ is finite, the classical perturbation analysis applies: see, for instance Kato [25], and all useful parameters of $\mathbf{\Pi}_1$ are analytic or Laurent in $\varepsilon = z^{*k}$, for ε sufficiently small.
2. $\mathsf{Pr}[T_k > n] \sim \mathsf{Pr}_1[T_k > n] + \mathcal{O}(e^{-C_{14}n\eta})$, where Pr_1 is related to $\mathbf{\Pi}_1$.
3. The parameters computed in Sec.2, A1-A3 have similar forms for the MC $\mathbf{\Pi}_1$. So we will now use all parameters corresponding expressions computed with $\mathbf{\Pi}_1$.
But we must now compare all these expressions with the corresponding ones related to $\mathbf{\Pi}$, which were used in the previous relations. For instance, let us compare $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_1$. We have

$$\begin{aligned}
(\boldsymbol{\pi}\mathbf{\Pi}^*)_i &= \pi(i) + \mu(i)\eta(1 + \mathcal{O}(\eta)) \quad , i \leq \kappa, \\
\boldsymbol{\pi}_1\mathbf{\Pi}_1 &= \boldsymbol{\pi}_1,
\end{aligned}$$

where $\boldsymbol{\mu} := C_{15}\left(\frac{z^{*i}}{1+z^{*i}}\right)$, (row vector). Hence

$$(\boldsymbol{\pi} - \boldsymbol{\pi}_1)\Pi_1 = \boldsymbol{\pi} - \boldsymbol{\pi}_1 - \boldsymbol{\pi}(\mathbf{\Pi}^* - \mathbf{\Pi}_1) + \boldsymbol{\mu}\eta(1 + \mathcal{O}(\eta)),$$

i.e.

$$(\boldsymbol{\pi} - \boldsymbol{\pi}_1)(i) = (\boldsymbol{\pi} - \boldsymbol{\pi}_1)\mathbf{1}\pi_1(i) + [\boldsymbol{\pi}(\mathbf{\Pi}^* - \mathbf{\Pi}_1)\mathbf{M}_1]_i - \eta(\boldsymbol{\mu}\mathbf{M}_1)_i(1 + \mathcal{O}(\eta)).$$

But

$$\boldsymbol{\pi}(\mathbf{\Pi}^* - \mathbf{\Pi}_1)_i = [i = \kappa]C_{16}\eta(1 + \mathcal{O}(\eta))$$

and $M_1(\cdot, i)$ is given by (45) (computed with $\mathbf{\Pi}_1$). Hence

$$(\boldsymbol{\pi} - \boldsymbol{\pi}_1)(i) = C_{17}\mathcal{O}(z^{*i}\eta).$$

Another useful vector is $\boldsymbol{\alpha}_1$, which is given by (see (18), (22))

$$\begin{aligned}
-(1 + z^{*l})[l \neq k] - (\mathbf{\Pi}^*\boldsymbol{\alpha})_l &= -(1 + z^{*l})C_{18}\eta(1 + \mathcal{O}(\eta)) - \gamma_1 - \alpha(l), \\
-(1 + z^{*l})[l \neq k] - (\mathbf{\Pi}_1\boldsymbol{\alpha}_1)_l &= -\gamma_1 - \alpha_1(l) - C_{19}\eta(1 + \mathcal{O}(\eta))
\end{aligned}$$

or

$$(\boldsymbol{\alpha}_1-\boldsymbol{\alpha})(l)-[\mathbf{\Pi}_1(\boldsymbol{\alpha}_1-\boldsymbol{\alpha})]_l=\eta[C_{18}(1+z^{*l})(1+\mathcal{O}(\eta))+C_{19}(1+\mathcal{O}(\eta))]+[(\mathbf{\Pi}_1-\mathbf{\Pi}^*)\boldsymbol{\alpha}]_l,$$

i.e.

$$\begin{aligned}\boldsymbol{\alpha}-\boldsymbol{\alpha}_1 &= C_{18}\eta\mathbf{M}_1(\mathbf{1}+\mathbf{z}^*)(1+\mathcal{O}(\eta))\\ &\quad+C_4[C_{12}\mathbf{M}_1^-(\mathbf{1}+\mathbf{z}^*)\eta(1+\mathcal{O}(\eta))+C_{13}M_1(\cdot,\kappa)\eta(1+\mathcal{O}(\eta))]\\ &\quad+\boldsymbol{\pi}_1(\boldsymbol{\alpha}_1-\boldsymbol{\alpha}),\end{aligned}$$

but $\boldsymbol{\pi}_1(\boldsymbol{\alpha}_1-\boldsymbol{\alpha})=\boldsymbol{\pi}_1\boldsymbol{\alpha}_1+(\boldsymbol{\pi}_1-\boldsymbol{\pi})\boldsymbol{\alpha}+\boldsymbol{\pi}\boldsymbol{\alpha}$, and after all algebra, the dominant term of $\boldsymbol{\alpha}_1-\boldsymbol{\alpha}$ is given by $C_{20}\eta(1+\mathcal{O}(\eta))$. Actually, all usefull parameters can be similarly compared and the relative difference is always given by a $\mathcal{O}(\eta)$. So we can safely use the expressions computed in Sec.2 and A1-A3 (with full matrix $\mathbf{\Pi}$) in all our relations.

A.5 Analytic analysis of M

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

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

But

$$\Pi(j,k)=\frac{1+z^{*j}}{z^{*j}}[z^{*j}z^{*k}][j\neq k]\frac{1}{1+z^{*k}},$$

similarly,

$$\Pi^2(j,k)=\frac{1+z^{*j}}{z^{*j}}[z^{*j}\sum_{l\neq j,k}z^{*l}z^{*k}]\frac{1}{1+z^{*k}}.$$

With (13), we obtain

$$S(j,k)=\lim_{w\to 1}\left[\frac{1+z^{*j}}{z^{*j}}[\theta^k][\phi(w,\theta,z^*|j)-wz^{*j}\theta^j]\frac{1}{1+z^{*k}}+\frac{w^2}{1-w}\frac{z^{*k}}{(1+z^{*k})^2h_w}\right]$$

or

$$\begin{aligned}S(j,k) &= \lim_{w\to 1}\left[\frac{1+z^{*j}}{z^{*j}}[\theta^k]\left[\frac{A_2(w,\theta,z^*|j)-wz^{*j}\theta^j}{w}\right.\right.\\ &+\left.\left.\frac{A_1(w,\theta,z^*)A_2(w,1,z^*|j)}{wh(w,z^*)}\right]\frac{1}{1+z^{*k}}+\frac{w}{1-w}\frac{z^{*k}}{(1+z^{*k})^2h_w}\right].\end{aligned}$$

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

$$\begin{aligned}S(j,k)=\lim_{\epsilon\to 0}&\left[\frac{z^{*k}}{(1+z^{*k})^2h_w\varepsilon}-\frac{z^{*k}}{(1+z^{*k})^2h_w}-\frac{z^{*2j}}{(1+z^{*j})z^{*j}}[j=k]\right.\\ &\left.+\frac{1+z^{*j}}{z^{*j}}\left[-\frac{\varphi_5(1)}{h_w\varepsilon}+\frac{\varphi_{5,w}(1)}{h_w}-\frac{\varphi_5(1)h_{ww}}{2h_w^2}\right]\frac{1}{1+z^{*k}}\right]\end{aligned}\quad(47)$$

where

$$\begin{aligned}
\varphi_5(w,\theta,j) &:= A_1(w,\theta,z^*)A_2(w,1,z^*|j)/w \\
\varphi_5(1) &:= [\theta^k]\varphi_5(w,\theta,j)\big|_{w=1} = \frac{z^{*j}z^{*k}}{(1+z^{*j})(1+z^{*k})} \\
\varphi_{5,w}(1) &:= [\theta^k]\varphi_{5,w}(w,\theta,j))\big|_{w=1} \\
&= -\frac{z^{*2j}z^{*k}}{(1+z^{*j})^2(1+z^{*k})} + \frac{z^{*j}z^{*k}}{(1+z^{*j})(1+z^{*k})^2}
\end{aligned}$$

so the singularity in (47) is removed. Also from [37], we know that $h_{ww} = 2C_{20}$, where

$$C_{20} = \sum_{k\geq 1} \frac{z^{*2k}}{(1+z^{*k})^3} = 1.63759377999796\ldots.$$

And finally,

$$\begin{aligned}
M(j,k) &= S(j,k) + [j=k] - \pi(k) \\
&= -\frac{z^{*j}}{(1+z^{*j}))}[j=k] - \frac{z^{*j}z^{*k}}{(1+z^{*j})(1+z^{*k})^2 h_w} + \frac{z^{*k}}{(1+z^{*k})^3 h_w} \\
&\quad + C_{21}\frac{z^{*k}}{(1+z^{*k})^2} + [j=k]
\end{aligned}$$

where $C_{21} := -C_{20}/h_w^2 = -.277746036541901\ldots$ from which we derive a better precision for C_4: $C_4 = C_{21} - 1 = -1.27746036541901\ldots$. All our first-order approximations of M in A.1-A.3 are now easily checked.