Average Case Analysis of Gosper's Algorithm for a Class of Urn Model Inputs

Olgica Milenkovic[1] and Kevin J. Compton[2]

Abstract. In this paper we perform an asymptotic average case analysis of some of the most important steps of Gosper's algorithm for indefinite summation of hypergeometric terms. The space of input functions of the algorithm is described in terms of urn models, and the analysis is performed by using specialized probabilistic transform techniques. We analyze two different types of urn model classes: one in which the input functions are assumed to be rational, and another for which a certain function of two inputs is assumed to be rational. The first set of results shows that the asymptotic complexity of the algorithm is the same within each of the two classes. The second set of results indicates that the complexity of the algorithm scales differently for the two classes of models: one can observe the logarithmic versus square root type of difference that is also present in other combinatorial models.

1. Introduction. Gosper's algorithm is an automatic procedure for evaluating sums of hypergeometric terms in the form of the difference of a hypergeometric term and a constant, provided such an expression exists. Since it often happens that during the analysis of a problem in combinatorial theory one encounters a large sum involving factorials and binomial coefficients, one would like to know whether or not that sum can be expressed in a simpler way. Answering this question without the help of a computer usually requires a thorough knowledge of existing combinatorial identities and sometimes very ingenious reasoning. Gosper's algorithm is a procedure that discovers the answer systematically. As for any other algorithm, it is important to have an assessment of the time requirements to perform certain steps of the procedure. To do so, it is necessary to set up a complete problem specification, including the probability measure over the acceptable inputs of the algorithm. Once the specifications are set the goal is to find expressions for the first moments of random time parameters of interest.

Despite the fact that Gosper's algorithm is one of the most important achievements in computer algebra, to date there are no results concerning the average running time of the algorithm [8], [26]. In this paper we investigate certain probabilistic and combinatorial problems arising in the analysis of Gosper's algorithm [10]. Although Gosper's algorithm itself is simple to describe and understand, the average time analysis is a very complicated task. The main problem is the complicated space of input functions of the algorithm. This

---

[1] Department of Electrical and Computer Engineering, University of Colorado, Boulder, CO 80309, USA. milenkov@schof.Colorado.edu.

[2] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA. kjc@eecs.umich.edu.

space consists of combinatorial functions and it is hard to define a simple but sufficiently general probability distribution for it that would allow for analyzing all the steps of the algorithm simultaneously. This is why we perform a partial average time analysis, focusing on one step in the procedure that has the largest influence on the running time of the algorithm and that may occur as a part of other algorithms as well. In order to perform the analysis, we use some newly developed probabilistic transforms described in a companion paper [20].

The paper is organized as follows. Section 2 presents a short description of Gosper's algorithm. Section 3 introduces the probabilistic models used for the input functions of the algorithm. Section 4 contains the derivations of the expected value of one of the most important variables in the algorithm. In Sections 5 and 6 the average case analysis of two distinct probabilistic input models is performed using the Poisson and negative-binomial transforms. Section 7 contains the concluding remarks and a short overview of future work.

Throughout the paper, $\mathcal{N}$ and $\mathcal{Z}$ stand for the set of natural numbers including zero and the set of integers, respectively, while $\mathcal{C}$ stands for the set of complex numbers; $a(x) \sim b(x)$ is a standard symbolic statement of the fact that $\lim_{x \to \infty} a(x)/b(x) = 1$.

2. A Short Description of Gosper's Algorithm. Assume that we are interested in evaluating sums of the form $S_n = \sum_{k=0}^{n-1} t_k$, where t_k is a hypergeometric term. We say that t_k is a hypergeometric term if $r(k) = t_{k+1}/t_k$ is a rational function of k. Throughout the paper we refer to $r(k)$ as the *hypergeometric ratio*. The indirect definition of t_k in terms of the hypergeometric ratio essentially implies that t_k is a constant multiple of a term of the form

(1) $$F([a_1, \ldots, a_m]; [b_1, \ldots, b_s], z)_k = \frac{a_1^{\underline{k}} \cdots a_m^{\underline{k}} \, z^k}{b_1^{\underline{k}} \cdots b_s^{\underline{k}} \, k!},$$

where $z \in \mathcal{C}$, and $a^{\underline{k}} = a(a-1) \cdots (a-k+1)$ denotes the falling factorial. Gosper [10], [22] designed an algorithm that can find a solution for S_n in terms of the difference of a hypergeometric term and a constant, or prove that no such formula exists.

The idea behind Gosper's procedure is based on the observation that the problem of finding closed forms for summations is analogous to the problem of finding closed forms for indefinite integrals. Therefore, evaluating S_n can be reduced to finding an "anti-derivative" z_k of the summation term t_k, i.e. a hypergeometric term z_k such that $z_{k+1} - z_k = t_k$. If such a term exists, then from the telescoping property it follows that

(2) $$S_n = \sum_{k=0}^{n-1} t_k = \sum_{k=0}^{n-1} (z_{k+1} - z_k) = \sum_{k=0}^{n-1} z_{k+1} - \sum_{k=0}^{n-1} z_k = z_n - z_0,$$

where z_0 is a constant. A sequence $\{t_k\}$ for which there exists another hypergeometric sequence $\{z_k\}$ satisfying (2) is called Gosper-summable. Without loss of generality, we assume from now on that $t_k \neq 0$, since otherwise the summation problem is trivial.

For a hypergeometric term z_k, the ratio

(3) $$y(k) = \frac{z_k}{t_k} = \frac{z_k}{z_{k+1} - z_k} = \frac{1}{z_{k+1}/z_k - 1}$$

is a rational function of k. Then $z_{k+1} - z_k = t_k$ implies that

$$(4) \qquad\qquad y(k+1)t_{k+1} - y(k)t_k = t_k.$$

Dividing both the left-hand and right-hand side of (4) by t_k, we obtain

$$y(k+1)\frac{t_{k+1}}{t_k} - y(k) = 1,$$

i.e.

$$(5) \qquad\qquad y(k+1)r(k) - y(k) = 1,$$

where $r(k) = t_{k+1}/t_k$ is the hypergeometric ratio. If there exists a rational function $y(k)$ satisfying (5), it can be obtained by using the following approach. Assume that $r(k) = f(k)/g(k)$, where $f(k)$ and $g(k)$ are relatively prime polynomials. If the greatest polynomial common divisor of $f(k)$ and $g(k+h)$, denoted $d_h(k) = \gcd(f(k), g(k+h))$, is a constant for all $h \in \mathcal{N}$, set $a(k) = f(k)$, $b(k) = g(k)$. Otherwise, form the set H of all values $h \geq 1$, $h \in \mathcal{N}$, for which $d_h(k) \neq \text{const}$. We say that h violates the gcd constraint if and only if $h \in H$. Next, take the smallest $h \in H$ and factor $r(k)$ as

$$(6) \qquad r(k) = \frac{f(k)/d_h(k)}{g(k)/d_h(k-h)} \frac{d_h(k)}{d_h(k-h)} = \frac{f_{(h)}(k)}{g_{(h)}(k)} \frac{d_h(k)}{d_h(k-h)},)$$

where

$$f_{(h)}(k) = \frac{f(k)}{d_h(k)} \quad \text{and} \quad g_{(h)}(k) = \frac{g(k)}{d_h(k)}.$$

The last term in (6) can be rewritten as

$$(7) \qquad \begin{aligned} \frac{d_h(k)}{d_h(k-h)} &= \frac{d_h(k)\, d_h(k-1) \cdots d_h(k-h+1)}{d_h(k-1)\, d_h(k-2) \cdots d_h(k-h)} = \frac{c(k+1)}{c(k)}, \\ c(k) &= d_h(k-1) \cdots d_h(k-h). \end{aligned}$$

Repeating the described procedure for all the remaining elements of H produces a sequence of functions $f_{(i)}, g_{(i)}, i \in H$, and an expression for the hypergeometric ratio of the form

$$(8) \qquad\qquad r(k) = \frac{a(k)}{b(k)} \frac{c(k+1)}{c(k)}.$$

The polynomials $a(k), b(k), c(k)$ are such that $\gcd(a(k), b(k+h))$ is a constant for all $h \in \mathcal{N}$. We refer to the above procedure as the *cancelling procedure*. The cancelling procedure can be performed for all rational functions $r(k)$ [22].

Since $y(k)$ is a rational function, one can write $y(k) = (b(k-1)x(k))/c(k)$, where $x(k)$ is another rational function of k. Plugging this expression for $y(k)$ and the expression for $r(k)$ given by (8) into (5) gives

$$(9) \qquad \begin{aligned} \frac{b(k)\, x(k+1)}{c(k+1)} r(k) - \frac{b(k-1)\, x(k)}{c(k)} &= 1 \quad \text{or} \\ a(k)x(k+1) - b(k-1)x(k) &= c(k). \end{aligned}$$

Provided $a(k), b(k), c(k)$ are constructed using the approach above, it can be shown that if there exists a rational function $x(k)$ satisfying (9), then this function is necessarily a polynomial. This is the key idea of Gosper's algorithm; its straightforward proof can be found in [22].

Once the polynomials $a(k), b(k)$ and $c(k)$ are determined, the next step in the algorithm is solving (9) for a polynomial function $x(k)$. Let the upper bound on the degree of $x(k)$ be $\leq d$, i.e. let $x(k) = \sum_{u=0}^{d} a_u k^u, a_d \neq 0$. Substituting this expression for $x(k)$ into (9) and solving the system of equations in the unknown coefficients a_m gives the desired solution for $x(k)$ and hence for $z(k)$ as well. If there is no solution for the system of equations, then t_k is not Gosper-summable. There are two different possibilities for the value of the degree d of $x(k)$:

- If $deg(a(k)) \neq deg(b(k))$ or $lc(a(k)) \neq lc(b(k))$, then

  (10) $$d = deg(c(k)) - \max\{deg(a(k)), deg(b(k))\},$$

  where $lc(a(k))$ denotes the leading coefficient of the polynomial $a(k)$.
- if $deg(a(k)) = deg(b(k))$ and $lc(a(k)) = lc(b(k)) = s$, then either

  (11) $$d = deg(c(k)) - deg(a(k)) + 1$$

  or

  $$d = (slc(b(k-1)) - slc(a(k)))/s,$$

  where $slc(a(k))$ denotes the second leading coefficient of the polynomial $a(k)$.

The last case occurs extremely rarely in practical problems [8], [22] and is not analyzed in this paper.

From the description of the algorithm, one can see that there are basically two parameters that influence the running time of the procedure. The first is the number of values of h that violate the gcd constraint. For each value of $h \in H$ one has to compute the gcd of two polynomials, perform several cancellations and rewrite the expressions for several polynomials. The second parameter is the set of degrees of the polynomials $a(k), b(k)$ and $c(k)$. The degrees of these polynomials determine the size of the system of linear equations to be solved in order to find a solution or prove that a solution does not exist (as we pointed out, the second formula for the degree of $x(k)$ that involves the second leading coefficients is of no practical use). Hence, in the forthcoming sections, we focus on evaluating the expected values of these parameters only.

EXAMPLE [22].   Assume we want to know if the sum

$$S_{n+1} = \sum_{k=0}^{n} (4k + 1) \frac{k!}{(2k + 1)!}$$

can be expressed in closed form. Since the summand $t_k = ((4k + 1) k!)/(2k + 1)!$ is a hypergeometric term, we can apply Gosper's algorithm in the following way. First,

$$r(k) = \frac{t_{k+1}}{t_k} = \frac{4k + 5}{2(4k + 1)(2k + 3)} = \frac{\left(k + \frac{5}{4}\right)}{4\left(k + \frac{1}{4}\right)\left(k + \frac{3}{2}\right)},$$

so that the roots of the linear terms in the numerator and denominator of $r(k)$ are $-\frac{5}{4}$, and $-\frac{1}{4}$ and $-\frac{3}{2}$, respectively. The difference of the roots $-\frac{5}{4}$ and $-\frac{3}{2}$ is not an integer, so for no value of $h \in \mathcal{N}$ will these terms cancel. Therefore, $2k+3$ is a factor of the polynomial $b(k)$. Similarly, one can see that $a(k) = 1$, and that $\gcd(k + \frac{5}{4}, k + h + \frac{1}{4}) = k + \frac{5}{4}$ for $h = 1$. Hence $c(k) = 4k + 1$. Now, (9) takes the form

$$(12) \qquad\qquad x(k+1) - 2(2k+1)x(k) = 4k + 1.$$

We notice that a polynomial of degree zero, namely $x(k) = -1$, is a solution of (12). This gives $z_k = (-2(2k+1)(4k+1)\,k!)/((4k+1)(2k+1)!) = (-2\,k!)/(2k)!$. Since the summation S_{n+1} ranges from zero to n rather than $n-1$, the closed form solution for the sum under consideration is

$$S_{n+1} = z_{n+1} - z_0 = 2 - \frac{n!}{(2n+1)!}.$$

3. Probabilistic Input Models. In this section we describe two probabilistic models for the summand t_k and the hypergeometric ratio $r(k)$ that capture most of the essence and generality of the space of hypergeometric terms and ratios, respectively.

3.1. *Probabilistic Model for t_k—the* **T** *Model.* In many combinatorial problems, t_k is a rational function that is already given in factored form as

$$(13) \qquad\qquad t_k = \frac{p(k)}{q(k)} = \prod_{i=1}^{n} \frac{(k - m_i)}{(k - l_i)},$$

where the m_i's and l_j's are either integers or rational numbers. The roots $\{m_i\}_{i=1}^n$ and $\{l_j\}_{j=1}^n$ can be partitioned into a set of sets $M_1, \ldots, M_u$ and $L_1, \ldots, L_u$, respectively, such that for any $m_i, m_j, \; i \neq j$, and $l_k, l_n, \; k \neq n$, the following properties hold:

- $m_i - m_j \in \mathcal{Z}$ if and only if $m_i, m_j \in M_s$, for some s;
- $m_i - l_k \in \mathcal{Z}$ if and only if $m_i \in M_t, \; l_i \in L_t$ for some t;
- $l_k - l_n \in \mathcal{Z}$ if and only if $l_k, l_n \in L_g$ for some g.

Under the assumptions above, each pair of sets $\{M_1, L_1\}, \{M_2, L_2\}, \ldots$ contains elements that differ by integer values, and elements within different pairs differ by non-integer values. Hence, t_k can be factored into rational functions with the roots of their numerators and denominators in $M_j, L_j, j = 1, \ldots, u$, respectively, and each such rational function can be analyzed separately. Hence, we assume from now on that m_i and l_i take integer values only. For the average case analysis it is of no importance what the signs of these integers are, as long as the range of values they take is bounded in the same way. We can therefore further restrict the values of m_i and l_j to positive integers only.

The choice for the distribution of the integer roots gives rise to two different probabilistic models. For the first model, which we refer to as the *uniform* **T** *model*, we assume that $M_i, \; L_i$ are two sequences of random integers drawn independently and uniformly with replacement from $\{1, \ldots, N\}$. For example, under this assumption, the sequences

of roots $(1, 1, 4, 4, 5, 6)$ and $(4, 1, 1, 6, 5, 4)$ are considered different, although they give rise to the same polynomial $(x - 1)^2(x - 4)^2(x - 5)(x - 6)$. It is clear that the **T** model is actually an urn model of the Maxwell–Boltzman type [6]. Under the Maxwell–Boltzman model, polynomials with roots of large multiplicity will be less likely than polynomials with roots of small multiplicity.

For the second model, which we refer to as the *multiset* **T** *model*, we assume that the sets M_i, L_i are two multisets chosen randomly from a set of equally likely multisets of size n over an alphabet of size N. The multiset model is an urn model of the Bose–Einstein type [6].

We note that for both the uniform and multiset model the polynomials in the numerator and denominator of the hypergeometric ratio are not necessarily co-prime. Hence, before proceeding with the analysis of the cancelling procedure, we have to examine the average number of cancellations needed to bring $f(k)$ and $g(k)$ into co-prime form. This number of cancellations determines the average value of the effective degree of both $f(k)$ and $g(k)$ used for further evaluations.

3.2. *Probabilistic Model for $r(k)$—the* **R** *Model*. Since all steps in the algorithm ultimately depend on the form of the hypergeometric ratio $r(k)$, one can consider introducing a probabilistic model for this rational function directly. The distribution of the roots of $f(k)$ and $g(k)$ as defined in (6) can be assumed to obey a uniform or multinomial distribution, similar to the case of the **T** model. In this model, which we refer to as the **R** model, there are no dependencies between the roots of $f(k)$ and $g(k)$, as opposed to the case of the **T** model.

3.3. *Advantages and Drawbacks of the Models*. The models proposed for t_k and the hypergeometric ratio $r(k)$ include some simplified assumptions that may not assure full generality of the model. First, the degree of t_k is usually a function of k and not fixed to a specified value of n, as we assumed. Second, the effective degree of t_k varies with the choices for the roots of $p(k)$ and $q(k)$. For example, the probability of choosing the same two sets of roots for $p(k)$ and $q(k)$ is not equal to zero, so that our models also include $t_k = 1$ in the analysis, although summands of this kind are of no practical interest. On the other hand, it is worth pointing out that there is no loss in generality in choosing the above models for t_k, as far as it concerns including factorial functions.

REMARK. The simplest assumption that can be made about t_k is that it is a polynomial, but in this case it is well known that the solution can be expressed using Bernoulli numbers [11]. Another case for which the solution is known is for $t(z) = p(z)/q(z)$, where $p(z)$ and $q(z)$ are polynomials such that $\deg q(z) \geq \deg p(z) + 2$, and $q(n) \neq 0$ for any $n \in \mathcal{Z}$. The solution is of the form [16]

$$(14) \qquad \sum_n r(n) = -\sum Res(r(z) \, t_1(z)), \qquad t_1(z) = \frac{\pi}{\sin(\pi z)},$$

where $Res(f)$ denotes the residuum [15] of f and the summation on the right-hand side ranges over all zeros of $q(z)$. We point out that the approach taken in this paper allows us to accommodate for different degrees of $p(k)$ and $q(k)$, but since this makes the notation

and the essential ideas harder to follow, we choose to restrict our attention only to the case where $p(k)$ and $q(k)$ have the same degree.

For clarity of exposition, in this paper we analyze in detail only the multiset **T** and **R** models. Detailed derivations for the other model can be found in the extended version of the paper, available at [27]. Some key results for the uniform model are briefly discussed in Section 7.

4. The Expected Value of the Degree of $a(k)$ and $c(k)$.

We observed in Section 2 that the degree of the polynomial $x(k)$, and therefore the running time of a linear equation solver, depends on the degree of the polynomials $a(k)$, $b(k)$ and $c(k)$. In this section we derive general formulas for the expected value of the degree of these polynomials that hold for both the **T** and **R** models as well as for both choices for the distribution of the roots.

4.1. *The* **T** *Model.* We introduce a uniform probability measure on the set of all sequences of length n and over the alphabet $1, \ldots, N$, or a uniform probability measure on the set of all multisets of size n and over the alphabet $1, \ldots, N$. In both cases the numbers determining the multiplicities of the roots in $p(k)$ and $q(k)$, respectively, become random variables. Let $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$, respectively, denote the above described random variables. Then

$$(15) \qquad t_k = \frac{p(k)}{q(k)} = \frac{(k-1)^{X_1}(k-2)^{X_2} \cdots (k-N)^{X_N}}{(k-1)^{Y_1}(k-2)^{Y_2} \cdots (k-N)^{Y_N}},$$

where

$$(16) \qquad \sum_{i=1}^{N} X_i = n, \qquad \sum_{i=1}^{N} Y_i = n.$$

For t_k given by (15) the hypergeometric ratio $r(k)$ can be written as

$$(17) \quad r(k) = \frac{t_{k+1}}{t_k} = \frac{f(k)}{g(k)} = \frac{k^{X_1}(k-1)^{X_2+Y_1} \cdots (k-(N-1))^{X_N+Y_{N-1}}(k-N)^{Y_N}}{k^{Y_1}(k-1)^{Y_2+X_1} \cdots (k-(N-1))^{Y_N+X_{N-1}}(k-N)^{X_N}}.$$

We use the following definitions/notation related to the **T** model throughout the paper:

$$(18) \qquad \begin{aligned} X_1^{(0)} &= X_1, Y_1^{(0)} &&= Y_1, \\ X_i^{(0)} &= X_i + Y_{i-1}, \quad Y_i^{(0)} = Y_i + X_{i-1}, \quad && i = 2, \ldots, N, \\ X_{N+1}^{(0)} &= Y_N, Y_{N+1}^{(0)} &&= X_N. \end{aligned}$$

4.2. *The* **R** *Model.* Similarly as for the **T** model, let $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$ be random variables counting the multiplicities of the roots in $\{1, \ldots, N\}$ of $f(k)$ and $g(k)$, respectively. In other words, let

$$(19) \qquad r(k) = \frac{f(k)}{g(k)} = \frac{(k-1)^{X_1}(k-2)^{X_2} \cdots (k-N)^{X_N}}{(k-1)^{Y_1}(k-2)^{Y_2} \cdots (k-N)^{Y_N}},$$

where

(20)
$$\sum_{i=1}^{N} X_i = n, \qquad \sum_{i=1}^{N} Y_i = n.$$

We use the following definitions related to the **R** model throughout the paper:

(21)
$$X_i^{(0)} = X_i, \qquad Y_i^{(0)} = Y_i, \qquad i = 1, \dots, N.$$

The change of variables in (18) and (21) allows us to simplify the notation for the multiplicities of the roots in the numerator (and denominator) of the hypergeometric ratio $r(k)$, and to preserve a unique notation, independent on the choice of the model. Notice that the numerator and denominator of $r(k)$ in the **R** model do not have roots equal to zero, i.e. there are only N options for the roots, while for the random **T** model there are effectively $N+1$ options for the roots. This difference in the number of symbols does not influence the asymptotic behavior of the algorithm, and is neglected. Additionally, the degree of the numerator and denominator of the hypergeometric ratio for the **T** model is $2n$, and the variables X_i, Y_i correspond to the root $i-1$.

The variables X_i, $i = 1, \dots, N$ (as well as Y_i, $i = 1, \dots, N$), in both the **T** and **R** models are dependent, due to the restriction imposed on them by (20). This fact makes a direct analysis of Gosper's algorithm extremely hard, if not impossible. In order to overcome this difficulty, we use probabilistic transform techniques developed in a companion paper [20]. The transform techniques described in [20] allow one to replace the dependent random variables X_i and Y_i by independent random variables with appropriate distributions. An exact asymptotic expressions for a function of X_i and Y_i can then be recovered by using the inverse transform, or by using specially designed Tauberian theorems, also presented in [20]. For the uniform and multiset model, the transforms that achieve the desired goal are the *Poisson* [9] and *negative-binomial* transform. A short overview of the properties of these transforms, as well the Tauberian inversion theorem used in this paper are presented in Appendix 1. The reader is advised to go through the exposition in Appendix 1 before proceeding to the next section.

4.3. *The Cancelling Procedure and deg($c(k)$).* Suppose that one has to perform the cancelling procedure on $f(k)$ and $g(k)$. The procedure changes the multiplicities of the roots of $f(k)$ and $g(k)$ as follows: first, in the pre-processing step we cancel common terms in $f(k)$ and $g(k)$, in order to make these polynomials co-prime. This leads to a decrease of the multiplicities $X_i^{(0)}$ and $Y_i^{(0)}$ to $X_i^{(1)}$ and $Y_i^{(1)}$, respectively, based on the rule

$$\begin{aligned}
X_i^{(1)} &= X_i^{(0)} - \min(Y_i^{(0)}, X_i^{(0)}) = \max(0, X_i^{(0)} - Y_i^{(0)}), \\
Y_i^{(1)} &= Y_i^{(0)} - \min(Y_i^{(0)}, X_i^{(0)}) = \max(0, Y_i^{(0)} - X_i^{(0)}).
\end{aligned}$$

The values of the new multiplicities $X_i^{(1)}$, $Y_i^{(1)}$, $i = 1, \dots, N+1$, represent the input of the cancelling procedure. For $h = 1$, the uncancelled roots of the polynomial $g(k)$, now corresponding to a polynomial $g_{(1)}(k)$, are shifted to the left by one. The common terms are cancelled from the polynomials $f_{(1)}(k)$ and $g_{(1)}(k)$ and adjoint to $c(k)$. This procedure is repeated for all other values of $h \leq N$. The shifting procedure can be

visualized as shown below:

$$
\begin{array}{cccccccc}
X_1^{(0)} & X_2^{(0)} & X_3^{(0)} & \cdots & X_N^{(0)} & X_{N+1}^{(0)} & f(k) \\
\hline
Y_1^{(0)} & Y_2^{(0)} & Y_3^{(0)} & \cdots & Y_N^{(0)} & Y_{N+1}^{(0)} & g(k) \downarrow \\
& Y_1^{(1)} & Y_2^{(1)} & Y_3^{(1)} & \cdots & Y_N^{(1)} & Y_{N+1}^{(1)} & g_{(1)}(k) \downarrow \\
Y_1^{(2)} & Y_2^{(2)} & Y_3^{(2)} & \cdots & Y_N^{(2)} & Y_{N+1}^{(2)} & g_{(2)}(k) \downarrow \\
\leftarrow & \downarrow & \downarrow & \downarrow & & &
\end{array}
$$

If $X_i^{(j)}$ and $Y_i^{(j)}$ represent the multiplicities of the ith root after the jth step, then $X_i^{(j)}$ and $Y_i^{(j)}$ obey the following recurrence:

(22)
$$
\begin{aligned}
X_i^{(j+1)} &= X_i^{(j)} - \min(Y_{i+j}^{(j)}, X_i^{(j)}) = \max(0, X_i^{(j)} - Y_{i+j}^{(j)}), \\
Y_i^{(j+1)} &= Y_i^{(j)} - \min(Y_i^{(j)}, X_{i-j}^{(j)}) = \max(0, Y_i^{(j)} - X_{i-j}^{(j)}) .
\end{aligned}
$$

If during the jth step of the procedure m terms in $f_{(h)}(k)$ and $g_{(h)}(k)$ are cancelled, then the degree of the polynomial $c(k)$ increases by jm. This implies that

(23)
$$
\begin{aligned}
deg(c(k)) &= \sum_{i=1}^{N} \sum_{j=1}^{N+1-i} j(X_i^{(j)} - X_i^{(j+1)}) \\
&= \sum_{j=1}^{N} \sum_{i=1}^{N+1-j} X_i^{(j)} - \sum_{j=1}^{N} (N+1-j) X_j^{(N+1-j)},
\end{aligned}
$$

where $\sum_{i=1}^{N+1-j} (X_i^{(j)} - X_i^{(j+1)})$ is the total number of cancellations for $h = j$ and where $X_i^{(j)} - X_i^{(j+1)} = 0$ for all $j \geq N+1-i$. For a fixed j, and both the **T** and **R** models, the random variables $X_i^{(j)}$, $2 \leq i \leq N$, are identically distributed and therefore $E[X_i^{(j)}] = E[X_2^{(j)}]$, $2 \leq i \leq N$.

Hence,

(24)
$$
\begin{aligned}
E[deg(c(k))] &= \sum_{j=1}^{N} \sum_{i=1}^{N+1-j} E[X_i^{(j)}] - \sum_{i=1}^{N} (N+1-i) E[X_i^{(N+1-i)}] \\
&= \sum_{j=1}^{N-1} E[X_1^{(j)}] + N \sum_{j=1}^{N-1} E[X_2^{(j)}] - 2 \sum_{j=1}^{N-1} j \, E[X_2^{(j)}].
\end{aligned}
$$

The last equation is a consequence of the fact that all roots equal to zero (one) in the **T** model (in the **R** model) are cancelled up to the $N-1$st step of the procedure, so that $E[X_1^{(N)}] = 0$. Since $\sum_{i=0}^{N+1} X_i^{(0)} = 2n$, the degree of $a(k)$ can be found in a similar way as

(25)
$$
\begin{aligned}
E[deg(a(k))] &= 2n - \sum_{i=1}^{N+1} \sum_{j=1}^{N} E[X_i^{(j-1)} - X_i^{(j)}] \\
&= \sum_{i=2}^{N+1} E[X_i^{(N+1-i)}].
\end{aligned}
$$

We show next that $X_l^{(j)}$, $j, l = 1, 2, \ldots$, can be expressed in terms of the values of $X_i^{(0)}$ and $Y_i^{(0)}$, $i = 1, \ldots, N$, only.

THEOREM 1.   *Let* $X_1^{(0)}, \ldots, X_N^{(0)}$ *and* $Y_1^{(0)}, \ldots, Y_N^{(0)}$ *be two arbitrary sets of positive integers. Define* $X_i^{(j+1)}$ *and* $Y_i^{(j+1)}$ *as in* (22). *Then for* $j \geq 1$ *it holds that*

(26)
$$X_i^{(j)} = \max \left(0, \min_{0 \leq k < j} \sum_{l=0}^{k} (X_{i+l}^{(0)} - Y_{i+l}^{(0)}) \right),$$

$$Y_i^{(j)} = \max \left(0, \min_{0 \leq k < j} \sum_{l=0}^{k} (Y_{i-l}^{(0)} - X_{i-l}^{(0)}) \right).$$

PROOF.   The proof follows by induction. Clearly, the statement is true for $j = 1$, based on (22). Assume the statement is true for all values of $j \leq J$. If $X_i^{(J)} = 0$ for some fixed i then $X_i^{(J+l)} = 0$ for $l \geq 1$, so that (26) holds automatically. Therefore, without loss of generality we may assume that $X_i^{(J)} > 0$. From (22) one has

(27)
$$X_i^{(J+1)} = \max(0, X_i^{(J)} - Y_{i+J}^{(J)}),$$

$$Y_i^{(J+1)} = \max(0, Y_i^{(J)} - X_{i-J}^{(J)}).$$

Define T_k and $\hat{T}_k$ as

(28)
$$T_k = \sum_{l=0}^{k} (X_{i+l}^{(0)} - Y_{i+l}^{(0)}),$$

$$\hat{T}_k = \sum_{l=0}^{k} (Y_{i+J-l}^{(0)} - X_{i+J-l}^{(0)}).$$

Then $\hat{T}_{J-k-1} = T_k - T_J$, so that $\min_{0 \leq k < J} \hat{T}_k = - \max_{0 \leq k < J} (T_J - T_k)$. Therefore, from the induction hypothesis it follows that

(29) $$X_i^{(J)} - Y_{i+J}^{(J)} = \max \left(0, \min_{0 \leq k < J} T_k \right) - \max \left(0, - \max_{0 \leq k < J} (T_J - T_k) \right)$$

$$= \max \left(0, \min_{0 \leq k < J} T_k \right) + \min \left(0, \max_{0 \leq k < J} (T_J - T_k) \right)$$

$$= \max \left(0, \min_{0 \leq k < J} T_k \right) + \min \left(0, T_J - \min_{0 \leq k < J} T_k \right)$$

$$= \max \left(0, \min_{0 \leq k < J} T_k \right) + \min \left(-T_J, - \min_{0 \leq k < J} T_k \right) + T_J$$

$$= \max \left(0, \min_{0 \leq k < J} T_k \right) - \max \left(T_J, \min_{0 \leq k < J} T_k \right) + T_J$$

$$= \max \left(T_J, T_J + \min_{0 \leq k < J} T_k \right) - \max \left(T_J, \min_{0 \leq k < J} T_k \right)$$

$$= \min_{0 \leq k < J+1} T_k.$$

The last equation is a consequence of the fact that $X_i^{(J)} > 0$. From the induction hypothesis it follows that (26) holds for $X_i^{(J)}$, so that $\min_{0 \le k < J} T_k > 0$. This, on the other hand, shows that $T_J + \min_{0 \le k < J} T_k > T_J$, and hence $\max(T_J, T_J + \min_{0 \le k < J} T_k) = T_J + \min_{0 \le k < J} T_k$. Now, if $T_J < \min_{0 \le k < J} T_k$, then the penultimate equation in (29) equals $T_J + \min_{0 \le k < J} T_k - \min_{0 \le k < J} T_k = T_J = \min_{0 \le k < J+1} T_k$. If $T_J \ge \min_{0 \le k < J} T_k$ then (29) equals $T_J + \min_{0 \le k < J} T_k - T_J = \min_{0 \le k < J} T_k = \min_{0 \le k < J+1} T_k$. This proves the desired formula for $X_i^{(J)}$, while the proof for $Y_i^{(J)}$ follows along the same lines. $\square$

COROLLARY 2.   *The random variables $X_i^{(j)}$ satisfy the following equations for the* **T** *model*:

$$X_1^{(j)} = \max\left(0, \min_{1 \le k \le j}(X_k - Y_k)\right),$$

$$X_i^{(j)} = \max\left(0, \min_{i \le k \le \min(N+1, j+i-1)}((X_k - Y_k) - (X_{i-1} - Y_{i-1}))\right) \quad \textit{for} \quad 2 \le i \le N;$$

and the following equation for the **R** *model*:

$$X_1^{(j)} = \max\left(0, \min_{1 \le k \le j}\left(\sum_{i=1}^{k}(X_i - Y_i)\right)\right).$$

PROOF.   From the definition of $X_i^{(0)}$ and $Y_i^{(0)}$ it follows that $\sum_{l=0}^{k}(X_{1+l}^{(0)} - Y_{1+l}^{(0)}) = X_{k+1} - Y_{k+1}$ and $\sum_{l=0}^{k}(X_{i+l}^{(0)} - Y_{i+l}^{(0)}) = X_{k+i} - Y_{k+i} - (X_i - Y_i)$ for $i \ge 2$, which gives the desired result for the **T** model. Notice that since the expressions for $X_1^{(j)}$ and $X_i^{(j)}$, $i \ge 2$, are different, the expected values of these variables are also different, as we already pointed it out in this section. The result for the **R** model follows from the same type of argument. $\square$

Based on the result of Theorem 1, in order to find the degree of $c(k)$ we need to compute the distribution of the minimum of a set of random variables, as given by Corollary 2. The problem connected with this evaluation is that the variables involved are dependent. For this reason, we use the transform techniques of [20] to achieve a straightforward transition from the case of dependent variables to the case of independent variables.

5. The T Model.   We apply next the transform techniques of [20] (see Appendix 1) to the problem of computing the expected values of the variables $X_i^{(j)}$ and $Y_i^{(j)}$, and therefore the expected value of $deg(c(k))$ as well. Throughout this section we use the following notation:

$$
\begin{aligned}
Z_l &= X_l - Y_l, \\
M_j &= \min(Z_1, \ldots, Z_j), \\
W_{j,l} &= \min(Z_l, \ldots, Z_{l+j-1}) - Z_{l-1}.
\end{aligned}
$$
(30)

Under the assumption that X_l and Y_l are independent random variables in the appropriate transform domain, the variables Z_l are independent and identically distributed.

For both the uniform and multiset **T** model, one has

$$P\{Z_1 = k\} = \sum_{s=0}^{\infty} P\{X_1 = s + k\} P\{Y_1 = s\},$$

$$P\{M_j \geq u\} = (P\{Z_1 \geq u\})^j,$$

$$E[X_1^{(j)}] = \sum_{u=1}^{\infty} P\{M_j \geq u\},$$

(31)

$$P\{M_j = u\} = P\{M_j \geq u\} - P\{M_j \geq u + 1\},$$

$$P\{W_{j,l} = u\} = \sum_{s=-\infty}^{\infty} P\{M_j = s + u\}\, P\{Z_{l-1} = s\},$$

$$E[X_2^{(j)}] = \sum_{u=1}^{\infty} u\, P\{W_{j,2} = u\}.$$

The second equation is a consequence of the fact that in the transform domain the random variables Z_i, $i = 1, \ldots, j$, are independent and identically distributed.

5.1. *The Multiset* **T** *Model.*   For the multiset **T** model, the appropriate probabilistic transform is the negative-binomial transform. In this case we start from the assumption that X_i and Y_i are independent, geometrically distributed random variables with parameter ρ. The expected values of M_j and $W_{j,l}$ in this case can be determined as follows. First, notice that for $s \geq 0$,

(32)
$$P\{Z_1 = s\} = \sum_{l=0}^{\infty} (1 - \rho)^2 \rho^{s+2l} = \frac{1 - \rho}{1 + \rho}\, \rho^s,$$

while for $s < 0$ it follows in a similar way that

(33)
$$P\{Z_1 = s\} = \frac{1 - \rho}{1 + \rho}\, \rho^{|s|}.$$

Therefore, for $s \geq 0$ one has

$$P\{Z_1 \geq s\} = \frac{1 - \rho}{1 + \rho} \sum_{l=s}^{\infty} \rho^l = \frac{\rho^s}{1 + \rho},$$

while for $s < 0$ one has

(34)
$$P\{Z_1 \geq s\} = \frac{1 - \rho}{1 + \rho} \sum_{l=s}^{\infty} \rho^{|l|} = \frac{1 + \rho - \rho^{|s|+1}}{1 + \rho}.$$

This implies that for $s \geq 0$,

(35)
$$P\{M_j \geq s\} = \frac{\rho^{sj}}{(1 + \rho)^j},$$

while for $s < 0$,

$$(36) \qquad P\{M_j \geq s\} = \left(\frac{1 + \rho - \rho^{|s|+1}}{1 + \rho}\right)^j.$$

From (35) and (36), one can deduce that

$$(37) \qquad \begin{aligned} P\{M_j = s\} &= \frac{\rho^{sj}(1 - \rho^j)}{(1 + \rho)^j} && \text{for} \quad s \geq 0, \\ P\{M_j = s\} &= \frac{(1 + \rho - \rho^{|s|+1})^j - (1 + \rho - \rho^{|s|})^j}{(1 + \rho)^j} && \text{for} \quad s < 0. \end{aligned}$$

Hence,

$$(38) \qquad E[X_1^{(j)}] = E[M_j; M_j > 0] = \sum_{s=1}^{\infty} \frac{\rho^{sj}}{(1 + \rho)^j} = \frac{\rho^j}{(1 + \rho)^j (1 - \rho^j)}.$$

For $j = 1$, the above formula becomes

$$(39) \qquad E[X_1^{(1)}] = \frac{\rho}{1 - \rho^2}.$$

In order to find $E[X_2^{(j)}]$ we proceed as follows. Let $k \geq 1$. Then

$$\begin{aligned} P\{W_{j,l} = k\} &= \sum_{u=-\infty}^{-k-1} P\{M_j = k + u\} P\{Z_1 = u\} + \sum_{u=-k}^{-1} P\{M_j = k + u\} P\{Z_1 = u\} \\ &+ \sum_{u=0}^{\infty} P\{M_j = k + u\} P\{Z_1 = u\}. \end{aligned}$$

Let the three summations in the above equation be denoted by Σ_1, Σ_2, Σ_3, respectively. After some tedious, but straightforward algebra it can be shown that Σ_1 takes the form $\Sigma_1 = C_{1,j}(\rho) \rho^k$, where

$$C_{1,j}(\rho) = \frac{(1 - \rho)^2}{1 + \rho} \sum_{u=0}^{\infty} \rho^u \left(1 - \frac{\rho^{u+1}}{1 + \rho}\right)^j - \frac{1 - \rho}{(1 + \rho)^{j+1}}.$$

Σ_2 can be found from (32) and (37) in a similar way to be $\Sigma_2 = C_{2,j}(\rho) (\rho^k - \rho^{kj})$, where

$$C_{2,j}(\rho) = \frac{(1 - \rho)(1 - \rho^j)}{(1 + \rho)^{j+1} (1 - \rho^{j-1})}.$$

Finally, $\Sigma_3 = C_{3,j}(\rho) \rho^{kj}$, with

$$C_{3,j}(\rho) = \frac{(1 - \rho)(1 - \rho^j)}{(1 + \rho)^{j+1} (1 - \rho^{j+1})}.$$

When combined, the equations above give

$$(40) \qquad E[X_2^{(j)}] = (C_{3,j}(\rho) - C_{2,j}(\rho)) \frac{\rho^j}{(1 - \rho^j)^2} + (C_{1,j}(\rho) + C_{2,j}(\rho)) \frac{\rho}{(1 - \rho)^2}.$$

After some algebraic manipulation the following expressions can be obtained:

$$(C_{3,j}(\rho) - C_{2,j}(\rho))\,\frac{\rho^j}{(1-\rho^j)^2} = -\frac{(1-\rho)^2\rho^{2j-1}}{(1+\rho)^j(1-\rho^{j-1})(1-\rho^j)(1-\rho^{j+1})}$$

and

$$(C_{1,j}(\rho) + C_{2,j}(\rho))\,\frac{\rho}{(1-\rho)^2} = \sum_{u=1}^{\infty}\frac{\rho^u}{1+\rho}\left(1-\frac{\rho^u}{1+\rho}\right)^j + \frac{\rho^j}{(1+\rho)^{j+1}(1-\rho^{j-1})}.$$

Therefore, $E[X_2^{(j)}]$ takes the form

$$(41) \quad E[X_2^{(j)}] = \sum_{u=1}^{\infty}\frac{\rho^u}{1+\rho}\left(1-\frac{\rho^u}{1+\rho}\right)^j + \frac{\rho^j(1-\rho^{j+2})}{(1+\rho)^{j+1}(1-\rho^j)(1-\rho^{j+1})}.$$

For $j = 1$ the above equation reduces to

$$(42) \quad E[X_2^{(1)}] = \frac{2\,\rho\,(1-\rho^3)}{(1-\rho)^2(1+\rho)^3}.$$

Notice that $E[X_1^{(0)}] - E[X_1^{(1)}]$ $(E[X_2^{(0)}] - E[X_2^{(1)}])$ represents the expected number of cancellations of the root equal to one (a root greater than one) in the pre-processing step of the cancelling procedure. These expected values are of interest because they show how many cancellations of the roots of $f(k)$ and $g(k)$, on average, are performed to bring these polynomials into co-prime form. Once this number of cancellations is known, one can deduce the effective degree of the polynomials in the numerator and denominator of $r(k)$.

Equations (39) and (42) can be used to prove the following result.

LEMMA 3.   *For large n and N, the effective degree of the polynomials $f(k)$ and $g(k)$ is of the form*

$$\frac{2n(n+N)}{(2n+N)^3}\,(3n^2 + 3nN + N^2).$$

PROOF.   The effective degree of $f(k)$ is asymptotically equal to $NE[X_2^{(1)}]$, and the result of the theorem is a consequence of the fact that Theorem 4 of [20] holds, and asymptotic inversion corresponds to substituting $n/(n+N)$ for ρ, as described in Appendix 1 (note that the degree of $f(k)$ and $g(k)$ in the **T** model is initially equal to $2n$).   □

Next, we establish an asymptotic formula for $E[X_2^{(j)}]$, based on (41). Let the first term in (41) be denoted by S_j. The formula for S_j is almost identical (except for the $1/(1 + \rho)$ term) to the formula for the expected value of a grade in the grading problem by Greene and Knuth [12].

The grading problem is concerned with a hypothetical situation of grading an exam with an infinite number of problems. With probability one grading is made a finite task by imposing the requirement that grading an exam stops at the first encounter of a wrong

solution. The grade A^+ is given to the student with the highest score, provided her/his score is unique. If there are several students sharing the highest score, no one receives an A^+. If we assume that there are n students in the class and each student gets the right answer with probability p, then it is straightforward to show that the expected value $E_{\max}$ of the maximum score in the class is

$$E_{\max} = \sum_{k=1}^{\infty}(1 - (1 - p^k)^n).$$

The expected value of the maximum score, provided that this score is unique, is proportional to a term of the form

(43)
$$\sum_{k=1}^{\infty} np^k(1 - p^k)^{n-1}.$$

The asymptotic behavior of $E_{\max}$ and the related quantity (43) can be determined by using two special techniques—the exponential and Gamma method—developed by Greene and Knuth [12] and De Bruijn [5]. An asymptotic formula for S_j can be deduced using the exponential method. For the sake of completeness, we sketch the idea underlying this method, referring the reader to p. 288 of [16] for more details and for the exact evaluations of the error term.

Let $x/j = \rho^u/(1 + \rho)$. Then, since $x/j < 1$ for $\rho < 1$, it follows that

$$\left(1 - \frac{\rho^u}{(1 + \rho)}\right)^j = \left(1 - \frac{x}{j}\right)^j \simeq e^{-x},$$

so that

(44)
$$S_j \simeq \frac{1}{j} \sum_{u=1}^{\infty} xe^{-x},$$

where x is a function of u. The last sum can be approximated by using the standard Γ-method [16]. The method is based on using the following representation of the exponential function in terms of the Gamma function:

(45)
$$e^{-x} = \int_{1/2} \Gamma(z) x^{-z} \, dz,$$

where $\int_{1/2}$ denotes a contour integral over a disc around the origin of radius $\frac{1}{2}$. Using (45), (44) becomes

(46)
$$\hat{S}_j = \frac{1}{j} \int_{1/2} \Gamma(z) \sum_{u=1}^{\infty} x^{1-z} \, dz = \frac{1}{j} \int_{1/2} \frac{\Gamma(z)}{(1 + \rho)^{1-z}} \frac{(j\rho)^{1-z}}{1 - \rho^{1-z}} \, dz.$$

Once $\hat{S}_j$ is expressed in the form of the integral (46), it is possible to use standard techniques in complex analysis to evaluate its asymptotic behavior, such as moving the contour of integration and estimating the integral from the values of the residues at the singularities. If $z_k = 1 + 2\pi ik/\log \rho$ are the roots of the denominator of the integrand in (44), it follows that

$$\hat{S}_j = \frac{1}{j}\left(\frac{1}{\log(\rho^{-1})} + \frac{4\pi}{\log^2 \rho}\sum_{k=2}^{\infty} \sqrt{k \log \rho}\, \csc(2\pi^2 k/\log \rho)\cos\left(\theta - \frac{2\pi k \log(j/(1+\rho))}{\log \rho}\right)\right),$$

with

$$\theta = -\alpha\gamma + \sum_{l \geq 1} \left(\frac{\alpha}{l} - \tan^{-1} \frac{\alpha}{l} \right), \qquad \alpha = \frac{2\pi k}{\log \rho},$$

and γ and csc denoting Euler's constant and the cosecant function, respectively. For fixed ρ the sum converges very fast to a value that is negligible compared with $1/\log(\rho^{-1})$. Since the error of approximating S_j by $\hat{S}_j$ is $o(1/j)$, and since this is also the order of the error of approximating $\hat{S}_j$ by $1/(\log(\rho^{-1}) j)$, it follows that $S_j = 1/(\log(\rho^{-1}) j) + o(1/j)$.

On the other hand, the second term in (41) can be bounded as

$$\frac{1 - \rho^3}{1 + \rho} \left(\frac{\rho}{1 + \rho} \right)^j \leq \frac{\rho^j (1 - \rho^{j+2})}{(1 + \rho)^{j+1}(1 - \rho^j)(1 - \rho^{j+1})} \leq \frac{1}{(1 - \rho^2)^2} \left(\frac{\rho}{1 + \rho} \right)^j,$$

which implies that

$$(47) \qquad S_j + \frac{1 - \rho^3}{1 + \rho} \left(\frac{\rho}{1 + \rho} \right)^j \leq E[X_2^{(j)}] \leq S_j + \frac{1}{(1 - \rho^2)^2} \left(\frac{\rho}{1 + \rho} \right)^j.$$

We recall the formula for $E[deg(c(k))]$ given in the previous section:

$$E[deg(c(k))] = \sum_{j=1}^{N-1} E[X_1^{(j)}] + N \sum_{j=1}^{N-1} E[X_2^{(j)}] - 2 \sum_{j=1}^{N-1} j E[X_2^{(j)}].$$

Denote the three summations in the above equation by E_1, E_2 and E_3, respectively. Then, based on (38), one has

$$(48) \qquad E_1 = \sum_{j=1}^{N-1} \frac{1}{1 - \rho^j} \left(\frac{\rho}{1 + \rho} \right)^j \leq \frac{1}{1 - \rho} \sum_{j=1}^{N-1} \left(\frac{\rho}{1 + \rho} \right)^j \leq \frac{\rho}{1 - \rho},$$

and similarly

$$E_1 \geq \frac{\rho}{1 - \rho^{N-1}} \left(1 - \left(\frac{\rho}{1 + \rho} \right)^{N-1} \right).$$

Since for $0 < \rho < 1$ one has $\rho/(1 + \rho) < 1$, for large values of N the last term in the previous equation is close to one. It therefore follows that for sufficiently large N, E_1 is only a function of ρ. From the expression for $E[X_2^{(j)}]$, and the well-known formula

$$\sum_{j=1}^{N} \frac{1}{j} = \log N + \gamma + O\left(\frac{1}{N} \right),$$

it follows that $E_2 = E_{2,1} + E_{2,2}$, where

$$(49) \qquad E_{2,1} = N \sum_{j=1}^{N-1} S_j \sim N \sum_{j=1}^{N-1} \frac{1}{\log(\rho^{-1}) j} \sim \frac{1}{\log(\rho^{-1})} N \log N$$

and

$$E_{2,2} \geq N \left(\frac{1-\rho^3}{1+\rho} \sum_{j=1}^{N-1} \left(\frac{\rho}{1+\rho} \right)^j \right) = N \left(\frac{1-\rho^3}{1+\rho} \rho \left(1 - \left(\frac{\rho}{1+\rho} \right)^{N-1} \right) \right).$$

Since $N \geq 2$, the bound above can be further simplified as

$$(50) \qquad E_{2,2} \geq N \frac{\rho(1-\rho^3)}{(1+\rho)^2},$$

and in a similar manner,

$$E_{2,2} \leq N \ \frac{1}{(1-\rho^2)^2} \sum_{j=1}^{N-1} \left(\frac{\rho}{1+\rho} \right)^j.$$

Hence,

$$(51) \qquad E_{2,2} \leq N \frac{\rho}{(1-\rho^2)^2}.$$

Equations (50) and (51) imply that $E_2 \sim N \log N / (\log(\rho^{-1}))$.

In order to find an asymptotic formula for E_3, we observe that

$$(52) \qquad \sum_{j=1}^{N-1} j x^j = \frac{x}{(1-x)^2} (1 - Nx^{N-1} + (N-1)x^N).$$

For $x = \rho/(1+\rho) < 1$ or $x = 1/(1+\rho) < 1$, the expression in (52) tends to $x/(1-x)^2$. Hence, the contribution to E_3 from the second term of the expression for $E[X_2^{(j)}]$ in (41) can be shown to be bounded by a function dependent only on ρ, and for large N negligible compared with $E_{2,1}$. Similarly, from the asymptotic estimate of the first term in (41), one can see that there exists a function of ρ, $C(\rho)$, such that

$$(53) \qquad \sum_{j=1}^{N-1} j \ S_j \sim C(\rho)N.$$

This is a consequence of the fact that $S_j \sim K/j$, for some constant K. The results presented in (41)–(53) suffice to prove the following theorem.

THEOREM 4.   *For the negative-binomial model, the expected value of the degree of $c(k)$ is asymptotically equal to*

$$\frac{N \log N}{\log(\rho^{-1})} \ .$$

Theorem 4 shows that the asymptotic expression for $E[deg(c(k))]$ is of the form of a product of a function of N and a function of ρ, analytic in any half-ring around the origin (the only singularity being at the origin itself). It is tedious, but straightforward, to verify that the second term in the asymptotic formula for $E[deg(c(k))]$ satisfies the conditions of Theorem 13 in Appendix 1. This implies the following result.

THEOREM 5. *The expected value of the degree of $c(k)$ for the multiset **T** model is asymptotically equal to*

$$E[deg(c(k))] \sim \frac{N \log N}{\log((n+N)/n)}.$$

Finally, the same type of arguments as presented in the previous section shows that $E[deg(a(k))]$ is of the order of $\log N$, and hence negligible compared with $E[deg(c(k))]$.

6. The R Model. Consider the random hypergeometric ratio model—the **R** model described in Section 3. Let S_k be the kth cumulative sum of the random variables $Z_1, Z_2, \ldots$ defined as in the previous section, and let $S_0 = 0$. Theorem 1 shows that in this case

$$X_1^{(j)} = \max(S_0, \min(S_1, \ldots, S_j))$$

for all $i = 1, \ldots, N-1$ and $X_i^{(j)} = X_1^{(j)} = 0$ for $j \geq N$, $i \geq 2$. As for the case of the first model, and for both the Poisson and negative-binomial distribution, the variables Z_i are independent and identically distributed. Hence, $S_k = \sum_{i=1}^{k} Z_i$ is effectively a random variable describing the position of a generalized random walk. There exists an extensive literature concerning properties of generalized random walks [25], [6], [7]. Of special interest to our analysis are results concerning the generating functions for the first moments of random walk variables. Among these is Spitzer's equation [25, p. 205], also referred to as Pollaszek–Spitzer's equation, establishing a formula for the generating function of the expected value of the maximum non-negative value of the random walk variable S_k. We use a similar approach to that given in [25] in order to compute the generating function of the mean value of the variable $\max(S_0, \min(S_1, \ldots, S_j))$. Due to more complicated initial conditions, the derivation is more involved than for the case considered by Spitzer.

THEOREM 6. *Let S_k be a generalized random walk, and define the random walk variable*

(54) $$m_j = \max(S_0, \min(S_1, \ldots, S_j)).$$

Then

(55) $$\sum_{j=0}^{\infty} E[m_j]t^j = \exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right) \sum_{k=1}^{\infty} \frac{t^k}{k} E[S_k; S_k > 0],$$

where $E[S_k; S_k > 0]$ is the expected value of S_k conditioned on $S_k > 0$.

REMARK. It is interesting to point out that the generating function of the expectation of the maximum $M_j = \max(S_0, S_1, \ldots, S_j)$ of the random walk is of a similar form, obtained by replacing $P\{S_k = 0\}$ in (55) by one. In this case one has

$$\exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k}\right) = \frac{1}{1-t},$$

and the analogue of (55) is of the form

$$\sum_{j=0}^{\infty} E[M_j] t^j = \frac{1}{1-t} \sum_{k=1}^{\infty} \frac{t^k}{k} E[S_k; S_k > 0].$$

This is Pollaszek–Spitzer's identity.

PROOF. Following the notation from [25], define the following stopping times:

$$\begin{aligned}
T^* &= \min(n, \ 1 \le n \le \infty : S_n < 0), \\
T' &= \min(n, \ 1 \le n \le \infty : S_n \le 0), \\
T_j &= \min(n, \ 0 \le n \le j : S_n = m_j),
\end{aligned}$$

as well as the event

(56) $$A_k = \{S_1 < 0, \ldots, S_{k-1} < 0, S_k > 0\}.$$

We first establish a more general result than given in the theorem, involving the generating function of $E[z^{S_j - m_j}]$ with respect to j. $E[z^{S_j - m_j}]$ can be evaluated by averaging it over the time T_j of the first occurrence of the value m_j in the sequence $S_0, S_1, \ldots, S_j$:

(57) $$\begin{aligned}
E[z^{S_j - m_j}] &= \sum_{k=0}^{j} E[z^{S_j - m_j}; T_j = k] \\
&= E[z^{S_j - m_j}; T_j = 0] + \sum_{k=1}^{j} E[z^{S_j - m_j}; T_j = k].
\end{aligned}$$

Assume first that $T_j > 0$. The event corresponding to $T_j = k$, $k > 0$, can be described by the following set of inequalities:

$$\begin{aligned}
S_k &> 0, \ S_1 > S_k, \ldots, S_{k-1} > S_k, \\
&S_{k+1} \ge S_k, \ldots, S_j \ge S_k.
\end{aligned}$$

This implies that $E[z^{S_j - m_j}; T_j = k]$ factors as

(58) $$\begin{aligned}
E[z^{S_j - m_j}; T_j = k] &= E[1, S_k > 0, S_1 > S_k, \ldots, S_{k-1} > S_k] \\
&\quad \times E[z^{S_j - S_k}, S_{k+1} \ge S_k, \ldots, S_j \ge S_k],
\end{aligned}$$

since the variables $S_k - S_1, \ldots, S_k - S_{k-1}$ depend only on $Z_1, \ldots, Z_k$, while the variables $S_{k+1} - S_k, \ldots, S_j - S_k$ depend only on $Z_{k+1}, \ldots, Z_j$. Now $S_k > S_0$, $S_k < S_1, \ldots, S_k < S_{k-1}$ implies that $Z_k + \cdots + Z_1 > 0$, $Z_k + \cdots + Z_2 < 0, \ldots, Z_k < 0$ and $S_{k+1} \ge S_k, \ldots, S_j \ge S_k$ implies that $Z_{k+1} \ge 0, \ldots, Z_{k+1} + \cdots + Z_j \ge 0$. Using (56), (58) becomes

(59) $$E[z^{S_j - m_j}; T_j = k] = P\{A_k\} E[z^{S_{j-k}}; T^* > j - k], \qquad k > 0.$$

Assume next that $T_j = 0$. In this case one must have $m_j = 0$ and at least one of the inequalities $S_i \leq 0, \ i \leq j$, has to hold (for otherwise $T_j > 0$), so that

$$E[z^{S_j - m_j}; T_j = 0] = E[z^{S_j}; S_1 \leq 0 \text{ or } S_2 \leq 0, \dots, \text{ or } S_j \leq 0]$$
$$= E[z^{S_j}; T' \leq j].$$

Since

$$E[z^{S_j}] = E[z^{S_j}; T' \leq j] + E[z^{S_j}; T' > j],$$

one has that

$$E[z^{S_j}; T' \leq j] = E[z^{S_j}] - E[z^{S_j}; T' > j].$$

Combining the above result with (59) gives

$$(60) \quad E[z^{S_j - m_j}] = E[z^{S_j}] - E[z^{S_j}; T' > j] + \sum_{k=1}^{j} P\{A_k\} E[z^{S_{j-k}}; T^* > j - k].$$

Multiplying both the left-hand and the right-hand side of (60) by t^j and summing over all $j \geq 0$ gives

$$(61) \quad \sum_{j=0}^{\infty} E[z^{S_j - m_j}] t^j = \left(\sum_{j=1}^{\infty} P\{A_j\} t^j \right) \left(\sum_{j=1}^{\infty} E[z^{S_j}; T^* > j] t^j \right)$$
$$+ \sum_{j=0}^{\infty} (E[z^{S_j}] - E[z^{S_j}; T' > j]) t^j,$$

or equivalently

$$\sum_{j=0}^{\infty} E[z^{S_j - m_j}] t^j = \beta(t) \gamma(t; z) + e(t; z) - e^*(t; z),$$

where

$$(62) \quad \begin{aligned} \beta(t) &= \sum_{j=1}^{\infty} P\{A_j\} t^j, \qquad \gamma(t; z) = \sum_{j=1}^{\infty} E[z^{S_j}; T^* > j] t^j, \\ e(t; z) &= \sum_{j=0}^{\infty} E[z^{S_j}] t^j \quad \text{and} \quad e^*(t; z) = \sum_{j=0}^{\infty} E[z^{S_j}; T' > j] t^j. \end{aligned}$$

Let us examine all the generating functions in (62) separately. Since

$$\begin{aligned} P\{A_j\} &= P\{S_1 < 0, \dots, S_{j-1} < 0, S_j > 0\} \\ &= P\{S_1 < 0, \dots, S_{j-1} < 0, S_j \geq 0\} - P\{S_1 < 0, \dots, S_{j-1} < 0, S_j = 0\}, \end{aligned}$$

it follows that $\beta(t) = \tau(t) - \alpha(t)$, where

$$\tau(t) = \sum_{j=1}^{\infty} P\{S_1 < 0, \dots, S_{j-1} < 0, S_j \geq 0\} t^j,$$

$$\alpha(t) = \sum_{j=1}^{\infty} P\{S_1 < 0, \dots, S_{j-1} < 0, S_j = 0\} t^j.$$

From p. 413 of [7], it is known that

$$\tau(t) = 1 - \exp\left(-\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k > 0\}\right),$$

$$\alpha(t) = 1 - \exp\left(-\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right).$$

Hence,

$$\beta(t) = \exp\left(-\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right) - \exp\left(-\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k > 0\}\right).$$

From p. 180 of [25], it is also known that

$$e^*(t; z) = \exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} E[z^{S_k}; S_k > 0]\right)$$

and

$$(63) \quad \gamma(t; z) = \exp\left(\sum_{k=1}^{\infty} E[z^{S_k}; S_k > 0]\right) \exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right)$$

$$= e^*(t; z) \exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right).$$

Substituting formula (63) into (6) gives

$$(64) \quad \sum_{j=0}^{\infty} E[z^{S_j - m_j}]t^j = \left(\beta(t) \exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right) - 1\right) e^*(t; z) + e(t; z).$$

Differentiating (64) with respect to z and setting $z = 1$ shows that

$$\sum_{j=0}^{\infty} E[S_j - m_j]t^j = \left(\beta(t) \exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right) - 1\right) (e^*(t; z))' \big|_{z=1} + \sum_{j=1}^{\infty} E[S_j]t^j.$$

Since S_j is a symmetric random variable, $E[S_j] = 0$ for all j, it follows that

$$-\sum_{j=0}^{\infty} E[m_j]t^j = \left(\beta(t) \exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right) - 1\right) (e^*(t; z))' \big|_{z=1}.$$

Observing that

$$\left(\beta(t) \exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right) - 1\right)$$

$$= -\exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right) \exp\left(-\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k > 0\}\right)$$

holds, and combining this equation with

$$\left(e^*(t; z)\right)' \big|_{z=1} = \exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k > 0\}\right) \sum_{k=1}^{\infty} \frac{t^k}{k} E[S_k; S_k > 0]$$

shows that

$$\sum_{j=0}^{\infty} E[m_j] t^j = \exp\left(\sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\}\right) \sum_{k=1}^{\infty} \frac{t^k}{k} E[S_k; S_k > 0].$$

This proves the claimed result. $\square$

We use the result of Theorem 6 to derive a formula for the asymptotic behavior of the expectations $E[X_i^{(j)}]$, $i \geq 1$, $j \geq 1$, for both the Poisson and negative-binomial models.

6.1. *The Multiset Model.*   In this section we derive an asymptotic expression for the coefficients of the generating function (55) for the negative-binomial model. We start by evaluating the probability $P\{S_k = 0\}$.

LEMMA 7.   *For $k \geq 1$, it holds that*

$$P\{S_k = 0\} = \frac{(1 - \rho^2)}{(1 + \rho)^{2k}} q_{k-1}(\rho^2),$$

where

$$q_{k-1}(x) = (1 - x)^k P_k\left(\frac{1 + x}{1 - x}\right),$$

and $P_k(x)$ is the kth Legendre polynomial [21].

PROOF.   We start by observing that S_k can be written in the form

(65) $$S_k = (X_1 + \cdots + X_k) - (Y_1 + \cdots + Y_k) = U_k - V_k,$$

where U_k, V_k are independent variables with a negative-binomial distribution with parameters (k, ρ), so that

$$P\{S_k = 0\} = \sum_{m=0}^{\infty} P\{U_k = m\} P\{V_k = m\} = \sum_{m=0}^{\infty} P\{U_k = m\}^2$$

$$= (1 - \rho)^{2k} \sum_{m=0}^{\infty} \binom{m + k - 1}{m}^2 \rho^{2m}.$$

Consider

$$G_j(x) = \sum_{m=0}^{\infty} \binom{m + j}{m}^2 x^m,$$

where j is an integer. From p. 16 of [24] one has

$$\binom{j+m}{m}^2 = \sum_{l=0}^{m} \binom{j}{l}^2 \binom{2j+m-l}{2j},$$

and it follows that $\binom{m+j}{m}^2$ is the convolution of two binomial terms. This implies that

$$G_j(x) = G_{1,j}(x) \, G_{2,j}(x),$$

where

(66)
$$G_{1,j}(x) = \sum_{m=0}^{\infty} \binom{2j+m}{2j} x^m = \frac{1}{(1-x)^{2j+1}},$$

$$G_{2,j}(x) = \sum_{m=0}^{j} \binom{j}{m}^2 x^m.$$

$G_{2,j}(x)$ ($q_j(x)$ in the notation of [24, p. 81]) is a polynomial closely connected to the Legendre polynomial $P_j(x)$ of order k [21, p. 12], defined as

$$P_j(x) = \frac{1}{2^j} \sum_{k=0}^{\lfloor j/2 \rfloor} (-1)^k \binom{2(j-k)}{j-k} \binom{j-k}{k} x^{j-2k}.$$

More specifically, one has

$$G_{2,j}(x) = q_j(x) = (1-x)^j P_j\left(\frac{1+x}{1-x}\right).$$

Now, setting $j = k - 1$ and $x = \rho^2$ in (66) and (6.1) gives

(67) $$P\{S_k = 0\} = (1-\rho)^{2k} \frac{q_{k-1}(\rho^2)}{(1-\rho^2)^{2k-1}} = \frac{(1-\rho^2)}{(1+\rho)^{2k}} \, q_{k-1}(\rho^2).$$

This proves the desired result.   □

The next result is concerned with a closed form for $L_\rho(t) = \sum_{k=1}^{\infty} P\{S_k = 0\} t^k / k$.

LEMMA 8.   *The function $L_\rho(t)$ is of the form*

$$L_\rho(t) = \log \left| \frac{1-\rho^2}{2\rho^2} \left(\left(\frac{1-\rho}{1+\rho}\right)^2 t^2 - 2\frac{1+\rho^2}{(1+\rho)^2} t + 1 \right) + \frac{(1-\rho)^2}{2\rho^2} t - \frac{1+\rho^2}{2\rho^2} \right|.$$

PROOF.   From (67) it follows that

(68) $$L_\rho(t) = \sum_{k=1}^{\infty} \frac{t^k}{k} P\{S_k = 0\} = \sum_{k=1}^{\infty} \frac{t^k}{k} \frac{(1-\rho^2)}{(1+\rho)^{2k}} q_{k-1}(\rho^2).$$

Let

$$\alpha(\gamma, x) \; = \; \sum_{k=1}^{\infty} \frac{\gamma^k}{k} \, q_{k-1}(x),$$

$$L_\rho(t) \; = \; (1 - x)\, \alpha(\gamma, x),$$

where $\gamma = t/(1 + \rho)^2$ and $x = \rho^2$. Then

$$\frac{d}{d\gamma} \alpha(\gamma, x) = \sum_{k=1}^{\infty} q_{k-1}(x)\gamma^{k-1} = \sum_{k=0}^{\infty} q_k(x)\gamma^k,$$

so that after expressing $q_k(x)$ in terms of the Legendre polynomial $P_k(x)$, one obtains the formula

$$\frac{d}{d\gamma} \alpha(\gamma, x) = \sum_{k=0}^{\infty} P_k \left(\frac{1 + x}{1 - x} \right) ((1 - x)\gamma)^k .$$

The generating function of the Legendre polynomials is of the well known form [21, p. 11]

(69)
$$\sum_{n=0}^{\infty} P_n(x)t^n = \left(1 - 2xt + t^2\right)^{-1/2},$$

where $(1 - 2xt + t^2)^{-1/2}$ denotes the particular branch which tends to one as t goes to zero. Hence, it follows that

$$\frac{d}{d\gamma} \alpha(\gamma, x) \; = \; \left(1 - 2\frac{1 + x}{1 - x}(1 - x)\gamma + (1 - x)^2\gamma^2\right)^{-1/2}$$

$$= \; \left(1 - 2(1 + x)\gamma + (1 - x)^2\gamma^2\right)^{-1/2}.$$

This implies that

$$\alpha(\gamma, x) = \int_0^\gamma \frac{dy}{\left(1 - 2(1 + x)y + (1 - x)^2\, y^2\right)^{1/2}},$$

since $\alpha(0, x) = 0$. Using the result on p. 13 of [1], which states that

$$\int \frac{dy}{(ay^2 + by + c)^{1/2}} = \frac{1}{\sqrt{a}} \, \log |2\sqrt{a}\, (ay^2 + by + c)^{1/2} + 2ay + b|,$$

for $a > 0$, one can show that

$$\int_0^\gamma \frac{dy}{(ay^2 + by + c)^{1/2}} = \frac{1}{\sqrt{a}} \left(\log |2\sqrt{a}(ay^2 + by + c)^{1/2} + 2ay + b| - \log |2\sqrt{ac} + b| \right).$$

Therefore,

$$\alpha(\gamma, x) = \frac{1}{1 - x} \, \log \left| \frac{1 - x}{2x}((1 - x)^2\gamma^2 - 2(1 + x)\gamma + 1)^{1/2} + \frac{(1 - x)^2}{2x}\gamma - \frac{1 + x}{2x} \right|,$$

since $|2\sqrt{ac} + b| = |2(1 - x) - 2(1 + x)| = 4x$ for the function under consideration. After reintroducing the expression for γ and x in terms of t and ρ, one gets

$$L_\rho(t) = \log\left| \frac{1 - \rho^2}{2\rho^2} \left(\left(\frac{1 - \rho}{1 + \rho}\right)^2 t^2 - 2\frac{1 + \rho^2}{(1 + \rho)^2} t + 1 \right) + \frac{(1 - \rho)^2}{2\rho^2} t - \frac{1 + \rho^2}{2\rho^2} \right|,$$

which is the desired result. □

From Lemma 8 one has that

$$\exp(L_\rho(t)) = \left| \frac{1 - \rho^2}{2\rho^2} \left(\left(\frac{1 - \rho}{1 + \rho}\right)^2 t^2 - 2\frac{1 + \rho^2}{(1 + \rho)^2} t + 1 \right)^{1/2} + \frac{(1 - \rho)^2}{2\rho^2} t - \frac{1 + \rho^2}{2\rho^2} \right|.$$

The expression within the absolute value of the above formula can easily be shown to be negative, so that the function $\exp(L_\rho(t))$ is of the form

$$\exp\left(L_\rho(t)\right) = \frac{1 + \rho^2}{2\rho^2} - \frac{(1 - \rho)^2}{2\rho^2} t - \frac{1 - \rho^2}{2\rho^2} \left(\left(\frac{1 - \rho}{1 + \rho}\right)^2 t^2 - 2\frac{1 + \rho^2}{(1 + \rho)^2} t + 1 \right)^{1/2}.$$

The function

$$\left(\left(\frac{1 - \rho}{1 + \rho}\right)^2 t^2 - 2\frac{1 + \rho^2}{(1 + \rho)^2} t + 1 \right)^{1/2}$$

is the generating function of the Gegenbauer (ultra-spherical) polynomials $C_j^{-1/2}(x)$ [23, p. 276], defined as

$$C_j^\omega(x) = \sum_{k=0}^{\lfloor j/2 \rfloor} \frac{(-1)^k (\omega)_{j-k} (2x)^{j-2k}}{k!\,(j - 2k)!},$$

where

$$(a)_j = a(a + 1) \cdots (a + j - 1).$$

LEMMA 9.   *The asymptotic expression for the coefficient*

(70) $$[t^j] \exp(L_\rho(t))$$

is of the form

(71) $$\frac{1 - \rho}{2\rho^2} \sqrt{\frac{\rho}{\pi}} \frac{1}{j^{3/2}} + O_\rho(j^{-1}),$$

where the constant in the O term depends on ρ.

PROOF.   The proof is a direct consequence of Darboux's method [15, p. 447]. For an alternative proof, using the relationship that exists between the above described ultraspherical polynomials and the derivative of Legendre polynomials, see [19]. □

The second generating function in the product of Theorem 6 cannot be determined in a form simple enough for direct application. On the other hand, the asymptotic behavior of the function $E[S_k; S_k > 0]/k$ can be determined from the Central Limit Theorem [3, p. 357]. This follows from the fact that S_k is a partial sum of independent random variables Z_i with finite mean and variance. Therefore, if S_k is asymptotically a Gaussian $N(0, \sigma)$ random variable, one can prove, based on p. 338 of [3], that

$$\frac{E[S_k; S_k > 0]}{k} \sim \frac{1}{k\sqrt{2\pi}\,\sigma} \int_0^\infty x \, \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = \frac{\sigma}{k\sqrt{2\pi}},$$

where $\sigma^2 = ks^2$, and $s^2 = 2\rho/(1-\rho)^2$ is the variance of Z_i (the difference of two geometrically distributed random variable with parameter ρ). Hence

(72) $$\frac{E[S_k; S_k > 0]}{k} \sim \frac{1}{1-\rho}\sqrt{\frac{\rho}{\pi k}}.$$

The approximation error of (72), which is important for later applications, can be found from the asymptotic Edgeworth expansion [7, p. 531]. Since the distribution of Z_i is symmetric around the origin (i.e. $P\{Z_k = u\} = P\{Z_k = -u\}$), it follows that all the odd order central moments of Z_i are equal to zero. In this case the Edgeworth expansion for the approximate asymptotic probability density function of S_k is of the form

$$\eta(x) + \eta(x)\left(\frac{1}{k}\frac{\mu_4 - \sigma^4}{24\,\sigma^4} H_4(x) + \cdots\right),$$

where $\eta(x)$ is the Gaussian probability density function and $H_4(x)$ is the Hermite polynomial of order four [21]. The moments of Z_i can be found in a straightforward manner to be

$$\sigma^2 = \frac{2\rho}{(1-\rho)^2}, \qquad \mu_4 = \frac{2r(1 + 14r + 17r^2 + 4r^3)}{(1-r)^4}.$$

Also, the numerical value of the integral

$$\frac{1}{\sqrt{2\pi}} \int_0^\infty x \, \exp\left(\frac{-x^2}{2}\right) H_4(x) \, dx$$

can be easily found to be 44. Hence, the asymptotic expansion of $E[S_k; S_k > 0]/k$ is of the form

$$\frac{1}{1-\rho}\sqrt{\frac{\rho}{\pi k}}\left(1 + \frac{1}{24k}\frac{1 + 13\rho + 17\rho^2 + 4\rho^3}{\rho} + \cdots\right).$$

One can show, using the definition of the cumulants [7, p. 531], that the functions depending on ρ, in the above expansion, are rational. For the mth term of the expansion, the numerator of this rational function is a polynomial of degree m, the denominator is $\rho^{m/2-1}$ and this rational function is multiplied by $1/k^{m/2-1}$.

Having derived the above asymptotic formulas, we proceed as follows. First, we establish an asymptotic formula for $E[m_j] = E[X_1^{(j)}]$, using the information provided

by the asymptotic formula for $E[S_k; S_k > 0]/k$, (6) and some specialized Tauberian theorems. Then we find an asymptotic expression for the expected value of the degree of $c(k)$. The error estimates derived in Lemma 9 and using the Edgeworth expansion will provide sufficient information for verifying that the transfer theorem conditions (described in connection with Theorem 6 in [20]) hold in this case. This will allow us to conclude that the asymptotic formula for the expected value of the degree of $c(k)$ has the same form in the transform and in the inverse transform domain.

THEOREM 10.   *For $E[m_j]$ specified by the generating function* (6), *one has the following asymptotic formula*:

$$E[m_j] \sim \frac{1}{\rho} \frac{E[S_j, S_j > 0]}{j} \sim \frac{1}{(1 - \rho) \sqrt{\rho \pi j}}.$$

PROOF.   Consider the two generating functions on the right-hand side of (6). First, both generating functions have non-negative coefficients, and the radius of convergence of these series can be determined from the ratio test. In both cases the radius of convergence is equal to one. Furthermore, the probability $P\{S_k = 0\}$ converges to zero at the rate of $1/\sqrt{k}$ (due to the property that S_k is asymptotically Gaussian). Therefore the series $\sum_k P\{S_k = 0\}/k$ converges.

The above discussion, together with the asymptotic estimates of Lemma 9 and (72), shows that the conditions of Wood's Tauberian theorem are satisfied (see Appendix 2). If

$$a_k = \frac{E[S_k; S_k > 0]}{k},$$
$$c(k) = [t^k] \exp(L_\rho(t)),$$

then

$$(73) \qquad E[m_j] = E[X_1^{(j)}] = \sum_{k=0}^{j-1} a_k c_{j-k} \sim a_j \sum_{k=0}^{\infty} c_k = \frac{1}{\rho (1 - \rho)} \sqrt{\frac{\rho}{\pi j}},$$

where the last equality follows from the fact that $\exp(L_\rho(1)) = 1/\rho$. This proves the claimed result. $\square$

In Appendix 2 we also compute the error term in the asymptotic formula given above. Notice that error terms based on Tauberian theorems are usually of poor quality, and that the convergence to the correct asymptotic is usually much faster than predicted by these theorems.

THEOREM 11.   *The expected value of the degree of $c(k)$ in the transform domain is asymptotically of the form*

$$\frac{2}{3(1 - \rho)\sqrt{\rho \pi}} N^{3/2}.$$

PROOF. From (24), it follows that

$$E[deg(c(k))] = \sum_{m=1}^{N-1} (N + 1 - 2m) E[X_1^{(m)}]$$

$$= (N + 1) \sum_{m=1}^{N-1} E[X_1^{(m)}] - 2 \sum_{m=1}^{N-1} m E[X_1^{(m)}].$$

Substituting (73) together with the error estimates derived in this section and Appendix 2 into the above equation gives

$$E[deg(c(k))] = (N + 1) \left(\frac{1}{(1 - \rho)\sqrt{\rho\,\pi}} \sum_{j=1}^{N-1} \frac{1}{\sqrt{j}} + O\left(\sum_{j=1}^{N-1} \frac{1}{j^{9/16}} \right) \right)$$

$$- 2 \left(\frac{1}{(1 - \rho)\sqrt{\rho\,\pi}} \sum_{j=1}^{N-1} \sqrt{j} + O\left(\sum_{j=1}^{N-1} j^{7/16} \right) \right),$$

so that after using the estimates for the summations in the above equations as given in Appendix 2, we obtain

$$E[deg(c(k))] = \frac{2N^{3/2}}{(1 - \rho)\sqrt{\rho\,\pi}} + O(N^{1+7/16}) - 2 \left(\frac{2N^{3/2}}{3(1 - \rho)\,\sqrt{\rho\,\pi}} + O(N^{1+7/16}) \right)$$

$$= \frac{2}{3(1 - \rho)\sqrt{\rho\,\pi}}\, N^{3/2} + O(N^{1+7/16}).$$

This proves the claimed result.   □

Theorem 11 combined with estimates of the dependence of the error term on ρ can be used to verify that the transfer theorem, and hence Theorem 8 as well, apply in this case. We do not present the details of these derivations in this paper, but we point out that this result is intuitively clear from the fact that the error terms are rational functions in ρ, with denominator ρ. In this case the dominant factor in the error term is of the form

$$\frac{1}{N} \frac{\binom{n+N}{n+1}}{\binom{n+N-1}{n}} = \frac{n + N}{(n + 1)\,N},$$

which for large n and N is negligible compared with the main term, independent of the relationship between the parameters n and N.

THEOREM 12. *The asymptotic formula for the expected value of the degree of $c(k)$ in the multiset* **R** *model is of the form*

$$\frac{2}{3\sqrt{\pi n}}\, (n + N)^{3/2}\, \sqrt{N}.$$

This theorem follows directly from Theorem 11 by setting $\rho = n/(n + N)$.

7. Summary of Results. In this paper we addressed the problem of average time analysis of Gosper's algorithm for indefinite summation. We defined several probabilistic models for the space of input functions that capture the most important characteristics of input functions encountered in practice. The proposed models represent special cases of some well known classical urn models. Based on the proposed models and some new transform techniques, we derived asymptotic expressions for the expected number of linear equations L_e that have to be solved in order to obtain a closed form solution for the sum. These expressions are of the following form:

- For the uniform **T** model, one can prove that in the transform domain the following asymptotic upper bound holds:

$$L_e \leq \frac{N \log N}{\log p(\lambda)} - \frac{2 N \log N}{\log \log N},$$

  where p is a function of λ that does not implicitly depend on N.
- For the multiset **T** model we proved that

$$(74) \qquad L_e \sim \frac{N \log N}{\log((n+N)/n)}$$

  provided that $n/(n+N) > 0.2$ holds.
- For the uniform **R** model, one can prove that for $n/N \geq 5$ it holds that

$$L_e \sim \frac{2}{3} \sqrt{\frac{n}{\pi}} \, \exp\left(\sqrt{\frac{N}{\pi n}} \frac{\zeta(\frac{3}{2})}{2}\right) N.$$

- For the multiset **R** model, we showed that

$$L_e \sim \frac{2}{3\sqrt{\pi n}} (n+N)^{3/2} \sqrt{N}.$$

Appendix 1. The probabilistic models for the distribution of the integer roots of t_k described in Section 3 are examples of classical urn models [6]. If the values of the roots are represented by the urns, numbered as $1, 2, \ldots, N$, then the balls distributed into the urns determine the multiplicity of each root. For example, if three balls are placed in urn number five, then the root equal to five has multiplicity three.

An urn model for which n distinct balls are distributed into N distinct urns randomly and equally likely is known as the uniform or Maxwell–Boltzman model. Clearly, this represents the uniform model for the distribution of the roots of t_k described in Section 3. For the uniform model, the joint distribution of the multiplicities of the roots $X_i, i = 1, \ldots, N$, is multinomial, and

$$P\{X_1 = m_1, \ldots, X_N = m_N\} = \frac{1}{N^n} \binom{n}{m_1, \ldots, m_2},$$

with $\sum_{i=1}^{N} m_i = n$. Another widely used model, known as the Bose–Einstein model, considers the balls to be indistinguishable, and the urns to be distinguishable. This statistic corresponds to the second model for the distribution of the roots of t_k described in Section 3. For this model, the joint distribution of the multiplicities of the roots can be described as

(75) $$P\{X_1 = m_1, \ldots, X_N = m_N\} = \frac{1}{\binom{n+N-1}{}},$$

for any choice of a multiset of size n formed over the alphabet $\{1, 2, \ldots, N\}$.

For both models, the variables X_i representing the multiplicities of each root are not independent, and the analysis of any function involving these variables is quite difficult. On the other hand, the dependence enters only through the requirement that the sum of the variables is fixed. In order to overcome this type of problem for the uniform model, one can assume that the multiplicities X_i are generated for each value of $i = 1, \ldots, N$ independently by a Poisson process with mean λ (where the choice of λ is described later). In order to obtain the correct result, one has to apply the inverse Poisson transform [16]. Assume that the quantity one wants to estimate for the algorithm under the Poisson assumption is $\gamma(\lambda, N)$ and under the multinomial model $g(n, N)$. Then

(76) $$\gamma(N, \lambda) = \sum_{k \geq 0} g(N, k) P\{X_1 + X_2 + \cdots + X_N = k\},$$

where $P\{X_1 + X_2 + \cdots + X_N = k\}$ is the probability that a sum of N independent Poisson variables with parameter λ, is equal to k. The sum of independent Poisson random variables is also a Poisson variable, with a parameter equal to the sum of the means of the individual variables. From (76) it follows that $g(n, N)$ can be recovered from $\gamma(\lambda, N)$ as

(77) $$g(n, N) = \frac{n!}{N^n} [\lambda^n](e^{N\lambda} \gamma(\lambda, N)).$$

For the second model of Section 3, the Poissonization method fails to produce the correct answer, since the distribution of the balls is not uniform. For this model, the correct solution can be obtained by using the negative-binomial transform, a transform introduced in a companion paper [20].

Assume that the multiplicity of each root in the multiset model is generated independently by a geometric distribution (i.e. a distribution with probability mass function $P\{X = k\} = (1-\rho)\rho^k$, where ρ is a parameter to be described later). Now, following the same approach as for the case of the Poisson transformation, we see that the probability mass function $P\{X_1 + X_2 + \cdots + X_N = k\} = \binom{n+N-1}{n}(1 - \rho)\rho^k$ is negative-binomial, since a sum of independent, geometrically distributed random variables is a variable with a negative-binomial distribution [20]. Hence, we can define a negative-binomial transform of the form

(78) $$\gamma(\rho, N) = \sum_{n \geq 0} \binom{n+N-1}{n} \rho^n (1 - \rho)^N g(n, N).$$

The inverse negative-binomial transform is of the form

(79) $$g(n, N) = \frac{1}{\binom{n+N-1}{n}} [\rho^n] \left(\frac{\gamma(\rho, N)}{(1 - \rho)^N} \right).$$

Based on (77) and (79) one can show that under certain conditions $g(n, N)$ and $\gamma(\lambda, N)$ (or correspondingly $\gamma(\rho, N)$) have the same asymptotic behavior for large n, and N a function of n. The conditions for the validity of this type of asymptotic result are summarized in the following theorem.

THEOREM 13. *Let $f(\lambda, N)$ be a functions that has a Taylor series with respect to λ, and for all N, given by*

$$(80) \qquad f(\lambda, N) = \sum_{k=0}^{\infty} F_k(N)\, \lambda^k,$$

where it is assumed that $F_k(N)$ is strictly positive for sufficiently large N, and let $\gamma(\lambda, \alpha(n))$ be a regular and single-valued function in some open ring $\mathcal{R} = \{R_1 < |\lambda| < R_2, 0 \le R_1 < R_2 < 1\}$. Additionally, assume that $\gamma(\lambda, \alpha(n))$ satisfies the following asymptotic equality:

$$(81) \qquad \gamma(\lambda, \alpha(n)) = \hat{\gamma}(\lambda, \alpha(n)) \left(1 + O\left(\frac{y(\lambda)}{h(n)}\right)\right),$$

where $\hat{\gamma}(\lambda, \alpha(n)) = c(n)b(\lambda)$, and $y(\lambda)$ and $b(\lambda)$ (which do not depend on n) are regular and single valued in the ring $\mathcal{R}$. Let $y(\lambda)b(\lambda)$ have a single dominant singularity. Assume furthermore that the coefficient Y_m of the Laurent expansion of $y(\lambda)b(\lambda)$ with respect to λ and the coefficient B_m of the Laurent expansion of $b(\lambda)$ have the property that $Y_m/(B_m h(n)) \to 0$, uniformly in $m \le n$, as $n \to \infty$. Let

$$(82) \qquad g(n, N) = \frac{1}{F_n(N)}\, [\lambda^n]\{f(\lambda, N)\gamma(\lambda, N)\},$$

and let N be a function of n, $\alpha(n)$, such that $\lim_{n\to\infty} F_{n-1}(\alpha(n))/F_n(\alpha(n))$ exists, and is equal to r. Furthermore, let $A_n < r$ be a sequence such that

$$(83) \qquad A_n \sim \frac{F_{n-1}(\alpha(n))}{F_n(\alpha(n))} \sim r.$$

Then

$$(84) \qquad g(n, \alpha(n)) \sim \hat{\gamma}\,(A_n,\, \alpha(n)),$$

provided the following conditions are satisfied:

(A1) *$b(\lambda)$ converges at $\lambda = A_n$ to a non-zero value.*

(A2) *For some constant C, and for both positive and negative integer values of m such that $|m| \le n$ and large n, it holds that*

$$(85) \qquad \left|\frac{F_{n-m}(\alpha(n))}{F_n(\alpha(n))}\right| \le C A_n^m.$$

The proof of this result is given in [20].

For the Poisson and negative-binomial transform $F_n(N) = N^n/n!$ and $F_n(N) = \binom{n+N-1}{n}$, respectively. Hence, one can show that condition (A2) is satisfied provided that we take $\lambda = A_n = n/N$ and $\rho = A_n = n/(n + N)$, respectively. The above theorem, together with the functions for λ and ρ described above, is used to determine the asymptotic behavior of the expected value of the degree of $c(k)$.

Appendix 2.   Tauberian theorems are partial converses of Abel's theorem [13], concerned with deducing the properties of sequences based on the properties of their weighted averages. In this work we are only concerned with Tauberian theorems that establish asymptotic results for the coefficients of the product of two generating functions, provided the asymptotic behavior of the coefficients of the constituent generating functions is known. A Tauberian theorem we applied several times to problems in this paper is Woods' Tauberian theorem [28], stated below.

THEOREM 14.   *Let $a(\lambda)$, $c(\lambda)$ be two power series of the form*

$$(86) \qquad a(\lambda) = \sum_{m \geq 0} a_m \, \lambda^m \quad and \quad c(\lambda) = \sum_{m \geq 0} c_m \, \lambda^m.$$

Assume that

$$\lim_{n \to \infty} \frac{a_{n-1}}{a_n} = r.$$

If

$$a_m \sim \frac{C \, \lambda^{-m}}{m^s} \quad and \quad c_m = O\left(\frac{\lambda^{-m}}{m}\right),$$

for some constant $C > 0$, and $s < 1$, and if $c(\lambda)$ converges absolutely at $\lambda = r$ to some non-zero value, then

$$b_m \sim c(\lambda) \, a_m.$$

We determine next the error term in the Woods Tauberian term, for the case of the generating functions given in Theorem 6. For this case, $c_m \sim m^{-3/2}$ and $a_m \sim m^{-1/2}$.

THEOREM 15.   *The error term in the asymptotic formula of Thereom 6, derived from Woods' Tauberian theorem is $O(m^{-9/16})$.*

PROOF.   From the derivation of Woods' Tauberian theorem [4], it follows that we have to estimate the following terms:

$$T_1 = \left| c(1) - \sum_{l \leq k} c_l \right|, \qquad T_2 = \sum_{l \geq 0} |c_l| \left| 1 - \frac{a_{k-\sqrt{k}}}{a_k} \right|, \qquad T_3 = \sum_{l \leq \sqrt{k}} |c_l|,$$

$$T_4 = O\left(\left(\frac{Q_k \, k}{M_k}\right)^\mu\right), \qquad T_5 = O\left(\left(\frac{M_k}{k}\right)^{1-\mu}\right),$$

where

$$Q_k = \sum_{l \geq \sqrt{k}} |c_l| \quad and \quad M_k = k\sqrt{Q_k} + \sqrt{k}.$$

In order to find estimates for $T_1, \ldots, T_5$, we need to recall some simple results from number theory, concerning the order of magnitude of certain simple summation

functions [2, p. 55]:

$$\sum_{k \leq x} \frac{1}{k^s} = \frac{x^{1-s}}{1-s} + \zeta(s) + O(x^{-s}), \qquad s > 0, \quad s \neq 1;$$

$$\sum_{k > x} \frac{1}{k^s} = O(x^{1-s}), \qquad s > 1;$$

$$\sum_{k \leq x} k^s = \frac{x^{1+s}}{1+s} + O(x^s), \qquad s > 0,$$

where $\zeta(x)$ is the Riemann Zeta-function. Based on the above formulas, the estimates for $T_1, \ldots, T_5$ become

$$T_1 = \sum_{l \geq k} \frac{1}{l^{3/2}} = O(k^{1-3/2}) = O\left(\frac{1}{\sqrt{k}}\right),$$

$$T_2 \leq \zeta\left(\frac{3}{2}\right) \left| 1 - \frac{\sqrt{k}}{\sqrt{k - \sqrt{k}}} \right| = C \left| 1 - \left(1 + \frac{\sqrt{k}}{k - \sqrt{k}}\right)^{1/2} \right|$$

$$\leq C\left(1 - 1 + \frac{\sqrt{k}}{k - \sqrt{k}}\right) = O\left(\frac{1}{\sqrt{k}}\right),$$

$$T_3 = \sum_{l \leq k} |c_l| - \sum_{l \leq \sqrt{k}} |c_l| = O\left(\frac{1}{\sqrt{k}}\right) + \zeta\left(\frac{3}{2}\right) - O\left(\frac{1}{k^{1/4}}\right) - \zeta\left(\frac{3}{2}\right) = O\left(\frac{1}{k^{1/4}}\right),$$

$$Q_k = \sum_{l > \sqrt{k}} |c_l| = O(k^{(1/2)(1-3/2)}) = O\left(\frac{1}{k^{1/4}}\right), \qquad M_k = \frac{k}{k^{1/8}} + \sqrt{k} = O(k^{7/8}),$$

and

$$T_4 = O\left(\frac{1}{k^{1/16}}\right), \qquad T_5 = O\left(\frac{1}{k^{1/16}}\right).$$

When combined, $T_1, \ldots, T_5$ give an error term for $E[m_k]$ (both the Poisson and negative-binomial models) of the order of $a_k / k^{1/16}$, which is $O(1/k^{9/16})$. □

LEMMA 16 [17]. *Consider two arbitrary sequences a_n and b_n, such that*

$$1 + \sum_{n=1}^{\infty} a_n x^n = \exp\left(\sum_{n=1}^{\infty} b_n x^n\right)$$

holds in a neighborhood of the origin.

- *If $|b_n| \leq K n^{-\alpha}$ for $n \geq 1$, and some $K > 0$ and $\alpha > 1$, then $|a_n| \leq e^{KL} n^{-\alpha}$, for $n \geq 1$, where $L = 2^{\alpha+1} \zeta(\alpha)$, and $\zeta(\alpha)$ is the Riemann Zeta-function.*
- *If $b_n = o\left(n^{-\alpha}\right)$ as $n \to \infty$, for some $\alpha > 1$ then $a_n = o\left(n^{-\alpha}\right)$.*

References

[1]   M. Abramowitz, I. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1964.

[2]   T. Apostol, *An Introduction to Analytic Number Theory*, Springer-Verlag, New York, 1976.

[3]   P. Bilingsley, *Probability and Measure*, Series in Probability and Mathematical Statistics, Wiley, New York, 1995.

[4]   S. Burris, *Density in Number Systems*, Mathematical Surveys and Monographs, vol. 86, American Mathematical Society, Providence, RI, 2001.

[5]   N. De Bruijn, *Asymptotic Methods in Analysis*, Dover, New York, 1981.

[6]   W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. I, Wiley, New York, 1967.

[7]   W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, Wiley, New York, 1967.

[8]   J. Gerhard, High Degree Solutions for Low Degree Equations, *Proc. ISSAC '98*, Rostock, pp. 284–289, 1998.

[9]   G.H. Gonnet, J.I. Munro, The Analysis of Linear Probing Sort by the Use of a New Mathematical Transform, *J. Algorithms*, vol. 5, pp. 451–470, 1984.

[10] R. W. Gosper Jr., Decision Procedures for Indefinite Hypergeometric Summation, *Proc. Nat. Acad. Sci. USA*, vol. 75, pp. 40–42, 1978.

[11] R. Graham, D. Knuth, O. Patashnik, *Concrete Mathematics*, Addison-Wesley, Reading, MA, 1994.

[12] D. Greene, D. Knuth, *Mathematics for the Analysis of Algorithms*, 2nd edn., Birkhauser, Boston, MA, 1982.

[13] G.H. Hardy, *Divergent Series*, Clarendon Press, Oxford, 1949.

[14] G.H. Hardy, J.E. Littlewood, G. Polya, *Inequalities*, Cambridge University Press, Cambridge, 1997.

[15] P. Henrici, *Applied and Combinatorial Complex Analysis*: *Special Functions–Integral Transforms–Asymptotics–Continued Fractions*, Wiley, New York, 1977.

[16] M. Hofri, *Analysis of Algorithms*: *Computational Methods and Mathematical Tools*, Oxford University Press, Oxford, 1995.

[17] A. Knopfmacher, J. Knopfmacher, Arithmetical Semigroups Related to Trees and Polyhedra, *J. Combin. Theory Ser. A*, vol. 86, pp. 85–102, 1991.

[18] D. Knuth, *The Art of Computer Programming*, Vols. 1–3, Addison-Wesley, Reading, MA, 1969.

[19] O. Milenkovic, Combinatorial Problems in Analysis of Algorithms and Coding Theory, Ph.D thesis, EECS Department, University of Michigan, Ann Arbor, MI, 2002.

[20] O. Milenkovic, K.J. Compton, Probabilistic Transforms for Combinatorial Urn Models, accepted for publication in *Combin. Probab. Comput.*

[21] D.S. Mitrinovic, *An Introduction to Special Functions*, Gradjevinska Knjiga, Belgrade, 1972.

[22] M. Petkovsek, H. Wilf, D. Zeilberger, $A = B$, A.K. Peters, Wellesley, MA, 1996.

[23] E. Rainville, *Special Functions*, MacMillan, New York, 1963.

[24] J. Riordan, *Combinatorial Identities*, Wiley, New York, 1968.

[25] F. Spitzer, *Principles of Random Walk*, Van Nostrand, Princeton, NJ, 1964.

[26] J. von zur Gathen, J. Gerhard, Fast Algorithms for Taylor Shifts and Certain Difference Equations, *Proc. ISSAC '97*, Maui, pp. 40–47, 1997.

[27] Web-page of O. Milenkovic, http://ece.colorado.edu/milenkov/.

[28] A. R. Woods, Counting Finite Models, *J. Symbolic Logic*, vol. 62, pp. 925–949, 1997.