Analysis of a Randomized Selection Algorithm Motivated by the LZ'77 Scheme

Mark Daniel Ward
Department of Mathematics
Purdue University
West Lafayette, IN 47907-2067
mward@math.purdue.edu

Wojciech Szpankowski*
Department of Computer Science
Purdue University
West Lafayette, IN 47907-2066
spa@cs.purdue.edu

Abstract

We consider a randomized selection algorithm that has n initial participants and a moderator. In each round of the process, each participant and the moderator throw a biased coin. Only the participants who throw the same result as the moderator stay in the game for subsequent rounds. With probability 1, all participants are eliminated in finitely many rounds. We let M_n denote the number of participants remaining in the game in the last nontrivial round. This simple algorithm has surprisingly many interesting applications. In particular, it models (asymptotically) the number of longest prefixes in the Lempel–Ziv '77 data compression scheme. Such multiplicity was used recently in [13] to design an error-resilient LZ'77 scheme.

We give precise asymptotic characteristics of the jth factorial moment of M_n for all $j \in \mathbb{N}$. Also, we present a detailed asymptotic description of the exponential generating function for M_n. In particular, we exhibit periodic fluctuation in the distribution of M_n, and we prove that no limiting distribution exists (however, we observe that the asymptotic distribution follows the *logarithmic series distribution* plus some fluctuations). The results we develop are proved by probabilistic and analytical techniques of the analysis of algorithms. In particular, we utilize recurrence relations, analytical poissonization and depoissonization, the Mellin transform, and complex analysis.

1 Introduction.

We consider a randomized selection algorithm that has n initial participants and a moderator. At the outset, n participants and one moderator are present. Each has a biased coin with probability p of showing heads when flipped, and we write $q = 1 - p$. At each stage of the selection process, the moderator flips its coin once; participants remain for subsequent rounds if and only if their result agrees with the moderator's result. Note that all participants are eliminated in finitely many rounds with probability 1. We let M_n denote the number of participants remaining in the last nontrivial round (i.e., the final round in which some participants still remain). Equivalent descriptions of the algorithm are given in the next section.

Briefly we explain the algorithm in terms of tries. Consider a trie built from strings of 0's and 1's drawn from an i.i.d. source. We restrict attention to the situation where n such strings have already been inserted into a trie. When the $(n + 1)$-st string is inserted into the trie, M_n denotes the size of the subtree that starts at the insertion point of this new string.

The results of our analysis of M_n yield information about the redundancy in the LZ'77 algorithm [18]. In LZ'77, for a given training sequence $X_1, \ldots, X_n$, the next phrase is the longest prefix of the uncompressed sequence $X_{n+1}, X_{n+2}, \ldots$ that occurs at least once in the training sequence $X_1, \ldots, X_n$. Such a phrase can be found by building a suffix tree from the training sequence and inserting the $(n + 1)$-st suffix into the tree. The depth of insertion is the length of the next LZ'77 phrase and the size of the subtree starting at the insertion point represents the number of potential phrases (i.e., any phrase can be chosen for encoding). The latter quantity is asymptotically equivalent to our M_n (constructed for independent tries) with an error bound of $O(\log n/n)$ (cf. [9, 16]). Finally, we observe that multiplicity of LZ'77 phrases is used in Lonardi and Szpankowski [13] to design an error resilient LZ'77 scheme called LZRS'77 (the "RS" denotes Reed-Solomon error-correcting coding). Thus precise analysis of M_n allows us to obtain detailed information about the redundancy of LZ'77 and its error resilient version LZRS'77.

Related problems have been studied. For instance, suppose the moderator is replaced; instead, participants remain in the selection process if and only if they throw heads. Also, if $M_n \neq 1$, then the selection process is deemed inconclusive and the entire selection process is repeated. Finally, H_n denotes the number of rounds until the selection process determines a conclusive "leader." Prodinger [14] first posed this problem and made a non-trivial analysis, but he considered fair (unbiased) coins. Then Fill *et. al.* [3] found the limiting distribution of the number of rounds, but they also utilized fair coins. Recently, Janson and Szpankowski [11] gave precise asymptotic information about $E[H_n]$,

*The research of this author was supported by NSF Grant CCR-0208709, and by NIH Grant R01 GM068959-01.

Var$[H_n]$, and the distribution of H_n; we note that the analysis in [11] dealt with biased coins.

A wealth of results have been published that are pertinent to the methodology developed below (see especially [4] and [17]). We strongly emphasize that these methods are widely applicable to a great variety of other problems. The precise asymptotic descriptions of the distribution of M_n and the factorial moments of M_n should entice others to continue utilizing such methods in studying related problems.

We establish the asymptotic distribution of M_n and the factorial moments of M_n. Note that a first order asymptotic solution for the distribution and the factorial moments in a cone about the positive x axis is not too difficult to obtain, but a second order asymptotic solution is relatively much more difficult to derive. Our method is to first *poissonize* the problem. In other words, we no longer require n to be fixed, but instead we let the number of initial participants in the selection process be a random variable that is Poisson distributed with mean n. Then we utilize the Mellin transform and analytic methods to obtain asymptotic solutions in the Poisson model. Finally, we *depoissonize* the results to obtain the asymptotic distribution and factorial moments of M_n, both accurate to second order. Interestingly, when $\ln p/\ln q$ is rational, we note that the asymptotic distribution and factorial moments of M_n exhibit fluctuations. Therefore M_n does not have a limiting distribution or limiting factorial moments, but we provide precise formulas for both quantities. In particular, we prove that the asymptotic distribution of M_n follows the *logarithmic series distribution* (plus some fluctuations), that is, $P(M_n = j) \approx (1/h)(p^j(1-p) + (1-p)^j p)/j$ where h is the entropy rate.

Our results are organized in the following way: In the next section, two theorems are given. Theorem 1 provides a precise asymptotic description of the distribution of M_n. Then Theorem 2 gives analogous results for the factorial moments of M_n. Both theorems contain results which are second order accurate. Then we briefly discuss the consequences of the two theorems. In particular, we elaborate on the fluctuations mentioned above. In the third section, we prove both theorems. As we just mentioned briefly, our methodology uses poissonization. So we utilize a depoissonization lemma of Jacquet and Szpankowski (see [10] and [17]).

2 Main Results.

We first give a mathematically rigorous formulation of the problem to be analyzed. Let p be fixed with $0 < p < 1$, and write $q = 1 - p$. Define $X(j)$ to be the string $X_1^{(j)} X_2^{(j)} X_3^{(j)} ...$, where $\{X_i^{(j)} \mid i, j \in \mathbb{N}\}$ is a collection of i.i.d. random variables on $\{0, 1\}$, with $P\{X_i^{(j)} = 0\} = p$. Let $l_j^{(n)} = \sup\{i \geq 0 \mid X_1^{(j)} \dots X_i^{(j)} = X_1^{(n+1)} \dots X_i^{(n+1)}\}$. In other words, when comparing the jth and $(n+1)$st strings, let $l_j^{(n)}$ denote the length of the longest common prefix of these two strings. Then define $L_n = \max_{j \leq n} l_j^{(n)}$. In other words, among the first n strings, let L_n denote the length of the longest common prefix with the $(n+1)$-st string. Finally, define $M_n = \#\{j \mid 1 \leq j \leq n,\ l_j^{(n)} = L_n\}$. So M_n is the number of the first n strings that have a common prefix of length L_n with the $(n+1)$-st string. By convention, let $M_0 = 0$. In passing we observe that if the strings $X(1), \dots, X(n)$ are suffixes of a single string, then our M_n is asymptotically equivalent to the multiplicity of phrases in the LZ'77 scheme.

Now we present the problem from the viewpoint of tries. The alignment $C_{j_1,\dots,j_k}$ among k strings $X(j_1), ..., X(j_k)$ is the length of the longest common prefix of the k strings. We observe that $l_j^{(n)} = C_{j,n+1}$. The kth depth $D_{n+1}(k)$ in a trie built over $n+1$ strings is the length of the path from the root of the trie to the leaf containing the kth string. Note $D_{n+1}(n+1) = \max\limits_{1 \leq j \leq n} C_{j,n+1} + 1$. Therefore $L_n = D_{n+1}(n+1) - 1$. Thus, in the context of tries, $M_n = \#\{j \mid 1 \leq j \leq n,\ C_{j,n+1} + 1 = D_{n+1}(n+1)\}$. That is, M_n is the size of a subtree starting at the branching point of a new insertion.

Define the exponential generating functions

$$G(z, u) = \sum_{n \geq 0} E[u^{M_n}] \frac{z^n}{n!}$$

$$W_j(z) = \sum_{n \geq 0} E[(M_n)^{\underline{j}}] \frac{z^n}{n!}$$

for complex $u \in \mathbb{C}$ and $j \in \mathbb{N}$. If $f : \mathbb{C} \to \mathbb{C}$, then the recurrence relation

$$E[f(M_n)] = p^n(qf(n) + pE[f(M_n)]) + q^n(pf(n) + qE[f(M_n)]) + \sum_{k=1}^{n-1} \binom{n}{k} p^k q^{n-k} \left(pE[f(M_k)] + qE[f(M_{n-k})]\right) \tag{2.1}$$

holds for all $n \in \mathbb{N}$. If $f(0) = 0$, then the recurrence also holds when $n = 0$. To verify (2.1), just consider the possible values of $X_1^{(j)}$ for $1 \leq j \leq n+1$. Two useful facts follow immediately from this recurrence relation. First, if $n \in \mathbb{N}$, then

$$E[u^{M_n}] = p^n(qu^n + pE[u^{M_n}]) + q^n(pu^n + qE[u^{M_n}]) + \sum_{k=1}^{n-1} \binom{n}{k} p^k q^{n-k} (pE[u^{M_k}] + qE[u^{M_{n-k}}]). \tag{2.2}$$

Also, if $j \in \mathbb{N}$ and $n \geq 0$ then

$$E[(M_n)^{\underline{j}}] = p^n(qn^{\underline{j}} + pE[(M_n)^{\underline{j}}]) + q^n(pn^{\underline{j}} + qE[(M_n)^{\underline{j}}]) + \sum_{k=1}^{n-1} \binom{n}{k} p^k q^{n-k} (pE[(M_k)^{\underline{j}}] + qE[(M_{n-k})^{\underline{j}}]). \tag{2.3}$$

We derive an asymptotic solution for these recurrence relations using poissonization, the Mellin transform, and depoissonization; details are given in the next section. These methods yield the following two theorems.

THEOREM 2.1. *Let* $z_k = \frac{2kr\pi i}{\ln p}$ $\forall k \in \mathbb{Z}$, *where* $\frac{\ln p}{\ln q} = \frac{r}{s}$ *for some relatively prime* $r, s \in \mathbb{Z}$ *(recall that we are interested in the situation where* $\frac{\ln p}{\ln q}$ *is rational). Then*

$$E[(M_n)^{\underline{j}}] = \Gamma(j)\frac{q(p/q)^j + p(q/p)^j}{h} + \delta_j(\log_{1/p} n) - \frac{1}{2}n\left(\frac{d^2}{dz^2}\delta_j(\log_{1/p} z)\right)\Bigg|_{z=n} + O(n^{-2})$$

where $\delta_j(t) =$

$$\sum_{k\neq 0} -\frac{e^{2kr\pi it}\Gamma(z_k+j)\left(p^j q^{-z_k-j+1} + q^j p^{-z_k-j+1}\right)}{p^{-z_k+1}\ln p + q^{-z_k+1}\ln q}$$

and Γ *is the Euler gamma function.*

Note that the term $-\frac{1}{2}n\left(\frac{d^2}{dz^2}\delta_j(\log_{1/p} z)\right)\Big|_{z=n}$ is $O(n^{-1})$. Note that δ_j is a periodic function that has small magnitude and exhibits fluctuation. For instance, when $p = 1/2$ then $|\delta_j(t)| \leq \frac{1}{\ln 2}\sum_{k\neq 0}\left|\Gamma\left(j - \frac{2ki\pi}{\ln 2}\right)\right|$. The approximate values of $\frac{1}{\ln 2}\sum_{k\neq 0}\left|\Gamma\left(j - \frac{2ki\pi}{\ln 2}\right)\right|$ are given below for the first ten values of j.

j	$\frac{1}{\ln 2}\sum_{k\neq 0}\left\|\Gamma\left(j - \frac{2ki\pi}{\ln 2}\right)\right\|$
1	1.4260×10^{-5}
2	1.3005×10^{-4}
3	1.2072×10^{-3}
4	1.1527×10^{-2}
5	1.1421×10^{-1}
6	1.1823×10^{0}
7	1.2853×10^{1}
8	1.4721×10^{2}
9	1.7798×10^{3}
10	2.2737×10^{4}

We note that, if $\ln p/\ln q$ is irrational, then $\delta_j(x) \to 0$ as $x \to \infty$. So δ_j does not exhibit fluctuation when $\ln p/\ln q$ is irrational.

The next result describes the asymptotic distribution of M_n.

THEOREM 2.2. *Let* $z_k = \frac{2kr\pi i}{\ln p}$ $\forall k \in \mathbb{Z}$, *where* $\frac{\ln p}{\ln q} = \frac{r}{s}$ *for some relatively prime* $r, s \in \mathbb{Z}$. *Then* $E[u^{M_n}] =$

$$E[u^{M_n}] = -\frac{q\ln(1-pu) + p\ln(1-qu)}{h} + \delta(\log_{1/p} n, u) - \frac{1}{2}n\left(\frac{\partial^2}{\partial z^2}\delta(\log_{1/p} z, u)\right)\Bigg|_{z=n} + O(n^{-2}) \tag{2.4}$$

where

$$\delta(t,u) = \sum_{k\neq 0} -e^{2kr\pi it}\Gamma(z_k) \times \frac{\left(q(1-pu)^{-z_k} + p(1-qu)^{-z_k} - p^{-z_k+1} - q^{-z_k+1}\right)}{p^{-z_k+1}\ln p + q^{-z_k+1}\ln q}.$$

and Γ *is the Euler gamma function. It follows immediately that*

$$E[u^{M_n}] = \sum_{j=1}^{\infty}\frac{p^j q + q^j p}{jh}u^j + \sum_{j=1}^{\infty}\sum_{k\neq 0} -\frac{e^{2kr\pi i\log_{1/p} n}\Gamma(z_k)(p^j q + q^j p)(z_k)^{\overline{j}}}{j!(p^{-z_k+1}\ln p + q^{-z_k+1}\ln q)}u^j + O(n^{-1}) \tag{2.5}$$

and

$$P(M_n = j) = \frac{p^j q + q^j p}{jh} + \sum_{k\neq 0} -\frac{e^{2kr\pi i\log_{1/p} n}\Gamma(z_k)(p^j q + q^j p)(z_k)^{\overline{j}}}{j!(p^{-z_k+1}\ln p + q^{-z_k+1}\ln q)} + O(n^{-1}) \tag{2.6}$$

Note that δ is a periodic function that has small magnitude and exhibits fluctuation. For instance, when $p = 1/2$ then $|\delta(t,u)| \leq \frac{2}{\ln 2}\sum_{k\neq 0}|\Gamma(-2ki\pi/\ln 2)| \approx 3.1463 \times 10^{-6}$. The non-fluctuating part of the distribution of $P(M_n = j)$ follows the *logarithmic series distribution*, as already mentioned above.

If $\ln p/\ln q$ is irrational and u is fixed, then we observe $\delta(x,u) \to 0$ as $x \to \infty$. Thus δ does not exhibit fluctuation when $\ln p/\ln q$ is irrational.

Remark: We emphasize that the same methodology can be used to obtain even more terms in the asymptotic formulae given in the two theorems.

3 Analysis and Proofs.

Now we present our analytical approach for proving Theorems 1 and 2. Our first strategy is to *poissonize* the problem. Then we utilize the Mellin transform and complex analysis; thus we obtain asymptotic descriptions of the distribution and factorial moments of M_n, but we emphasize that these results are valid for the *poissonized* model of the problem. We must depoissonize our results in order to find the asymptotic distribution and factorial moments of M_n in the original model.

3.1 Poissonization. We first utilize analytical poissonization. The idea is to replace the fixed-size population model (i.e., the model in which the number of initial participants n in the selection process is fixed) by a poissonized model in which the number of initial participants is a Poisson random variable with mean n. This is affectionately referred to as "poissonizing" the problem. So we let the number of initial participants in the selection process be N, a random variable that has Poisson distribution and mean n (i.e., $P(N = j) = e^{-n}n^j/j!\ \forall j \geq 0$). We apply the Poisson transform to the exponential generating functions $G(z,u)$ and $W_j(z)$, which yields:

$$\widetilde{G}(z,u) = \sum_{n\geq 0} E[u^{M_n}]\frac{z^n}{n!}e^{-z}$$

and

$$\widetilde{W}_j(z) = \sum_{n\geq 0} E[(M_n)^{\underline{j}}]\frac{z^n}{n!}e^{-z}.$$

By using (2.2) to expand the coefficients of z^n in $G(z,u)$ for $n \geq 1$, we observe $G(z,u) =$

$$qe^{puz} + pe^{quz} - pe^{qz} - qe^{pz} + pG(pz,u)e^{qz} + qG(qz,u)e^{pz}. \quad (3.7)$$

Similarly, we apply (2.3) to the coefficients of z^n in $W_j(z)$ for $n \geq 1$ to see that $W_j(z) =$

$$q(pz)^j e^{pz} + p(qz)^j e^{qz} + pW_j(pz)e^{qz} + qW_j(qz)e^{pz} \quad (3.8)$$

for all $j \in \mathbb{N}$.

We observe that $\widetilde{G}(z,u) = G(z,u)e^{-z}$. If we multiply by e^{-z} throughout (3.7) and then simplify, we obtain

$$\begin{aligned}\widetilde{G}(z,u) &= qe^{(pu-1)z} + pe^{(qu-1)z} - pe^{-pz} - qe^{-qz} \\ &\quad + p\widetilde{G}(pz,u) + q\widetilde{G}(qz,u).\end{aligned} \quad (3.9)$$

Similarly, from (3.8) we know that if $j \in \mathbb{N}$ then

$$\widetilde{W}_j(z) = q(pz)^j e^{-qz} + p(qz)^j e^{-pz} + p\widetilde{W}_j(pz) + q\widetilde{W}_j(qz). \quad (3.10)$$

Note that the functional equations (3.9) and (3.10) for the poissonized versions of $G(z,u)$ and $W_j(z)$ are simpler than the corresponding equations (3.7) and (3.8) from the original (Bernoulli) model. We solve (3.9) and (3.10) asymptotically for large $z \in \mathbb{R}$.

3.2 Mellin Transform. If f is a complex-valued function which is continuous on $(0,\infty)$ and is locally integrable, then the Mellin transform of f is defined as

$$\mathcal{M}[f(x);s] = f^*(s) = \int_0^\infty f(x)x^{s-1}\,dx$$

(see [5] and page 400 of [17]). Three basic properties of the Mellin transform are useful in proving the next two results. We observe that

$$\begin{aligned}\mathcal{M}[x^j f(x);s] &= \int_0^\infty x^j f(x)x^{s-1}\,dx \\ &= \int_0^\infty f(x)x^{s+j-1}\,dx \\ &= f^*(s+j) \\ &= \mathcal{M}[f(x);s+j].\end{aligned}$$

If $\mu > 0$ we also notice

$$\begin{aligned}\mathcal{M}[f(\mu x);s] &= \int_0^\infty f(\mu x)x^{s-1}\,dx \\ &= \mu^{-s}\int_0^\infty f(x)x^{s-1}\,dx \\ &= \mu^{-s}f^*(s) = \mu^{-s}\mathcal{M}[f(x);s].\end{aligned}$$

Also

$$\mathcal{M}[e^{-x};s] = \int_0^\infty e^{-x}x^{-s}\,dx = \Gamma(s).$$

We first find the fundamental strip of $\widetilde{G}(x,u)$. By (3.9), we observe that

$$\begin{aligned}\widetilde{G}(x,u) &= qe^{(pu-1)x} + pe^{(qu-1)x} - pe^{-px} - qe^{-qx} \\ &\quad + p\widetilde{G}(px,u) + q\widetilde{G}(qx,u) \\ &= q\sum_{k\geq 0}\frac{((pu-1)x)^k}{k!} + p\sum_{k\geq 0}\frac{((qu-1)x)^k}{k!} \\ &\quad - p\sum_{k\geq 0}\frac{(-px)^k}{k!} - q\sum_{k\geq 0}\frac{(-qx)^k}{k!} \\ &\quad + p\sum_{n\geq 0}E[u^{M_n}]\frac{(px)^n}{n!}\sum_{k\geq 0}\frac{(-px)^k}{k!} \\ &\quad + q\sum_{n\geq 0}E[u^{M_n}]\frac{(qx)^n}{n!}\sum_{k\geq 0}\frac{(-qx)^k}{k!} \\ &\to q + q(pu-1)x + p + p(qu-1)x \\ &\quad - p + p^2x - q + q^2x \\ &\quad + p + p^2ux - p^2x + q + q^2ux - q^2x \\ &\quad\quad \text{as } x \to 0 \\ &= (u-1)x + 1\end{aligned}$$

We notice that $\widetilde{G}(x,u) \to 1$ as $x \to 0$, but we want to instead have $\widetilde{G}(x,u) = O(x)$ as $x \to 0$. So we replace $\widetilde{G}(x,u)$ by writing $\widehat{G}(x,u) = \widetilde{G}(x,u) - 1$. We expect $\widetilde{G}(x,u) = O(1) = O(x^0)$ as $x \to \infty$. Therefore the fundamental strip of $\widehat{G}(x,u)$ includes $(-1,0)$.

We next determine the fundamental strip of $\widetilde{W}_j(x)$. By (3.10), we know

$$\begin{aligned}
\widetilde{W}_j(x) &= q(px)^j e^{-qx} + p(qx)^j e^{-px} \\
&\quad + p\widetilde{W}_j(px) + q\widetilde{W}_j(qx) \\
&= q(px)^j \sum_{k\geq 0} \frac{(-qx)^k}{k!} + p(qx)^j \sum_{k\geq 0} \frac{(-px)^k}{k!} \\
&\quad + p \sum_{n\geq 0} E[(M_n)^{\underline{j}}] \frac{(px)^n}{n!} \sum_{k\geq 0} \frac{(-px)^k}{k!} \\
&\quad + q \sum_{n\geq 0} E[(M_n)^{\underline{j}}] \frac{(qx)^n}{n!} \sum_{k\geq 0} \frac{(-qx)^k}{k!} \\
&\to q(px)^j + p(qx)^j \\
&\quad + pE[(M_j)^{\underline{j}}]\frac{(px)^j}{j!} + qE[(M_j)^{\underline{j}}]\frac{(qx)^j}{j!} \\
&\quad \text{as } x \to 0 \\
&\quad \text{since } E[(M_n)^{\underline{j}}] = 0 \;\; \forall\, 0 \leq n < j \\
&= O(x^j)
\end{aligned}$$

We expect $\widetilde{W}_j(x) = O(1) = O(x^0)$ as $x \to \infty$. So $(-j, 0)$ is the fundamental strip of $\widetilde{W}_j(x)$.

If $u \in \mathbb{R}$ with $u < \min\{1/p, 1/q\}$ and if $\mathrm{Re}(s) \in (-1, 0)$ then it follows from (3.9) and the properties of the Mellin transform given above that $\widehat{G}^*(s,u) =$

$$\frac{\Gamma(s)\left(q(1-pu)^{-s} + p(1-qu)^{-s} - p^{-s+1} - q^{-s+1}\right)}{1 - p^{-s+1} - q^{-s+1}}.$$

If $j \in \mathbb{N}$ and $\mathrm{Re}(s) \in (-j, 0)$, then by (3.10) and the properties of the Mellin transform we mentioned, we see that

$$\widetilde{W}_j^*(s) = \frac{\Gamma(s+j)\left(p^j q^{-s-j+1} + q^j p^{-s-j+1}\right)}{1 - p^{-s+1} - q^{-s+1}}.$$

We note that the Mellin transform is a special case of the Fourier transform. So there is an inverse Mellin transform. Since $\widetilde{W}_j$ is continuous on $(0,\infty)$, then

$$\widetilde{W}_j(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \widetilde{W}_j^*(s) x^{-s}\, ds$$

if $c \in (-\alpha, -\beta)$, where $(-\alpha, -\beta)$ is the fundamental strip of $\widetilde{W}_j$. Thus

$$\widetilde{W}_j(x) = \frac{1}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \widetilde{W}_j^*(s) x^{-s}\, ds$$

since $c = -1/2$ is in the fundamental strip of $\widetilde{W}_j(x)$ $\forall j \in \mathbb{N}$.

Similarly

$$\widehat{G}(x,u) = \frac{1}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \widehat{G}^*(s,u) x^{-s}\, ds$$

since $c = -1/2$ is in the fundamental strip of $\widehat{G}(x,u)$.

3.3 Results for the Poisson Model. We are restricting attention to the case where $\ln p/\ln q$ is rational. Thus we can write $\ln p/\ln q = r/t$ for some relatively prime $r, t \in \mathbb{Z}$. Then, by a theorem of Jacquet and Schachinger (see page 356 of [17]), we know that the set of poles of $\widetilde{W}_j^*(s)x^{-s}$ is exactly $\left\{z_k = \frac{2kr\pi i}{\ln p} \mid k \in \mathbb{Z}\right\}$. We also observe that $\widetilde{W}_j^*(s)x^{-s}$ has simple poles at each z_k. Now we assume that $u \neq 1$. Then $\widehat{G}^*(s,u)x^{-s}$ has the same set of poles as $\widetilde{W}_j^*(s)x^{-s}$, each of which is a simple pole.

Let T_1 denote the line segment from $-\frac{1}{2} - iA$ to $-\frac{1}{2} + iA$ in the complex plane, where A is a large real number. Let T_2 denote the line segment from $-\frac{1}{2} + iA$ to $M + iA$. Let T_3 denote the line segment from $M + iA$ to $M - iA$. Let T_4 denote the line segment from $M - iA$ to $-\frac{1}{2} - iA$. Now we claim that, if $j \in \mathbb{N}$ and $z_k = \frac{2kr\pi i}{\ln p}$, then

$$\widetilde{W}_j(x) = \sum_{k\in\mathbb{Z}} -\mathrm{Res}[\widetilde{W}_j^*(s)x^{-s};\; z_k] + O(x^{-M}). \tag{3.11}$$

Using the Cauchy residue theorem [1], integrating clockwise around the curve described by T_1, T_2, T_3, T_4, we have

$$\begin{aligned}
\widetilde{W}_j(x) &= \frac{1}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \widetilde{W}_j^*(s) x^{-s}\, ds \\
&= \lim_{A\to\infty} \frac{1}{2\pi i}\int_{T_1} \widetilde{W}_j^*(s) x^{-s}\, ds \\
&= \lim_{A\to\infty} \Bigg(\sum -\mathrm{Res}[\widetilde{W}_j^*(s)x^{-s};\; z = a_l] \\
&\quad - \frac{1}{2\pi i}\left(\int_{T_2} + \int_{T_3} + \int_{T_4}\right) \widetilde{W}_j^*(s) x^{-s}\, ds \Bigg)
\end{aligned}$$

where the sum is taken over all poles a_l of $\widetilde{W}_j^*(s)x^{-s}$ in the region bounded by T_1, T_2, T_3, T_4.

By the smallness property of the Mellin transform (see page 402 of [17]), we observe that

$$\frac{1}{2\pi i}\left(\int_{T_2} + \int_{T_4}\right) \widetilde{W}_j^*(s) x^{-s}\, ds = O(A^{-1}).$$

We also observe (see page 408 of [17]) that

$$\begin{aligned}
&\left|\frac{1}{2\pi i}\int_{T_3} \widetilde{W}_j^*(s) x^{-s}\, ds\right| \\
&= \left|\frac{1}{2\pi i}\int_{M+i\infty}^{M-i\infty} \widetilde{W}_j^*(s) x^{-s}\, ds\right| \\
&= \left|\frac{1}{2\pi i}\int_{\infty}^{-\infty} \widetilde{W}_j^*(M+it) x^{-M-it}\, ds\right| \\
&\leq \left|x^{-M}/2\pi\right| \int_{\infty}^{-\infty} \left|\widetilde{W}_j^*(M+it)\right|\, ds
\end{aligned}$$

$$= \quad O(x^{-M}).$$

Combining these results proves the claim made in (3.11).

The same reasoning shows that

$$\widehat{G}(x,u) = \sum_{k\in\mathbb{Z}} -\mathrm{Res}[\widehat{G}^*(s,u)x^{-s};\ z_k] + O(x^{-M}). \tag{3.12}$$

We make the observation that

$$\begin{aligned} x^{-z_k} = x^{-2kr\pi i/\ln p} &= \exp \ln x^{2kr\pi i/\ln(1/p)} \\ &= \exp\left(\ln x^{2kr\pi i}/\ln(1/p)\right) \\ &= e^{2kr\pi i \log_{1/p} x}. \end{aligned}$$

Using this observation, we claim that if $j \in \mathbb{N}$ then

$$\widetilde{W}_j(x) = \Gamma(j)\frac{q(p/q)^j + p(q/p)^j}{h} + \delta_j(\log_{1/p} x) + O(x^{-M}) \tag{3.13}$$

where $h = -p\ln p - q\ln q$ denotes entropy and where $\delta_j(t) =$

$$\sum_{k\neq 0} -\frac{e^{2kr\pi i t}\Gamma(z_k+j)\left(p^j q^{-z_k-j+1} + q^j p^{-z_k-j+1}\right)}{p^{-z_k+1}\ln p + q^{-z_k+1}\ln q}.$$

To prove the claim, we first observe that, if $k \in \mathbb{Z}$, then

$$\begin{aligned} &\mathrm{Res}[\widetilde{W}_j^*(s)x^{-s};\ z_k] \\ &= x^{-z_k}\mathrm{Res}[\widetilde{W}_j^*(s);\ z_k] \\ &= e^{2kr\pi i \log_{1/p} x}\Gamma(z_k+j) \\ &\quad \times \frac{\left(p^j q^{-z_k-j+1} + q^j p^{-z_k-j+1}\right)}{p^{-z_k+1}\ln p + q^{-z_k+1}\ln q}. \end{aligned}$$

Now the claim made in (3.13) follows immediately from (3.11).

Now we claim that

$$\begin{aligned} \widehat{G}(x,u) &= -\frac{q\ln(1-pu) + p\ln(1-qu)}{h} \\ &\quad - 1 + \delta(\log_{1/p} x, u) + O(x^{-M}) \end{aligned} \tag{3.14}$$

where $h = -p\ln p - q\ln q$ denotes entropy and where $\delta(t,u) =$

$$\sum_{k\neq 0} -e^{2kr\pi i t}\Gamma(z_k) \times \frac{\left(q(1-pu)^{-z_k} + p(1-qu)^{-z_k} - p^{-z_k+1} - q^{-z_k+1}\right)}{p^{-z_k+1}\ln p + q^{-z_k+1}\ln q}.$$

The proof is similar to the proof of (3.13). If $k \neq 0$ then

$$\begin{aligned} &\mathrm{Res}[\widehat{G}^*(s,u)x^{-s};\ z_k] \\ &= x^{-z_k}\mathrm{Res}[\widehat{G}^*(s,u);\ z_k] \\ &= e^{2kr\pi i \log_{1/p} x} \\ &\quad \times \frac{\Gamma(z_k)}{p^{-z_k+1}\ln p + q^{-z_k+1}\ln q} \\ &\quad \times \left(q(1-pu)^{-z_k} + p(1-qu)^{-z_k} \right. \\ &\qquad \left. -p^{-z_k+1} - q^{-z_k+1}\right) \end{aligned}$$

Now we compute $\mathrm{Res}[\widehat{G}^*(s,u)x^{-s};\ z_0]$. We first observe that

$$\begin{aligned} &\Gamma(s)\left(q(1-pu)^{-s} + p(1-qu)^{-s} - p^{-s+1} - q^{-s+1}\right) \\ &= (s^{-1} + O(1))(-q\ln(1-pu)s - p\ln(1-qu)s \\ &\qquad + p\ln(p)s + q\ln(q)s + O(s^2)) \\ &= -q\ln(1-pu) - p\ln(1-qu) \\ &\quad + p\ln p + q\ln q + O(s). \end{aligned}$$

It follows that

$$\begin{aligned} \mathrm{Res}[\widehat{G}^*(s,u)x^{-s};\ z_0] &= x^{-z_0}\mathrm{Res}[\widehat{G}^*(s,u);\ z_0] \\ &= \frac{q\ln(1-pu) + p\ln(1-qu)}{h} \\ &\quad + 1. \end{aligned}$$

Combining these results, the claim given in (3.14) now follows from (3.12).

As an immediate corollary of (3.14), we see that

$$\begin{aligned} \widetilde{G}(x,u) &= -\frac{q\ln(1-pu) + p\ln(1-qu)}{h} \\ &\quad + \delta(\log_{1/p} x, u) + O(x^{-M}). \end{aligned}$$

We note that, if $\ln p/\ln q$ is irrational and u is fixed, then $\delta_j(x) \to 0$ and $\delta(x,u) \to 0$ as $x \to \infty$. Thus δ_j and δ do not exhibit fluctuation when $\ln p/\ln q$ is irrational.

3.4 Depoissonization. Recall that, in the original problem statement, n is a large, fixed integer. Most of our analysis has utilized a model where n is a Poisson random variable. Therefore, to obtain results about the problem we originally stated, it is necessary to depoissonize our results. We utilize the depoissonization techniques discussed in [10] and Chapter 10 of [17], especially the Depoissonization Lemma, to prove Theorems 1 and 2.

For the reader convenience we recall here some depoissonization results of [10]. Recall that a measurable function $\psi\colon (0,\infty) \to (0,\infty)$ is *slowly varying* if $\psi(tx)/\psi(x) \to 1$ as $x \to \infty$ for every fixed $t > 0$.

THEOREM 3.1. *Assume that $\widetilde{G}(z) = \sum_{n=0}^{\infty} g_n \frac{z^n}{n!}e^{-z}$ is a Poisson transform of a sequence g_n which is an entire function of a complex variable z. Suppose that there exist real constants $a < 1$, β, $\theta \in (0,\pi/2)$, c_1, c_2, and z_0, and a slowly varying function ψ such that the following conditions hold, where S_θ is the cone $S_\theta = \{z:\ |\arg(z)| \leq \theta\}$:*

(I) *For all $z \in S_\theta$ with $|z| \geq z_0$,*

$$|\widetilde{G}(z)| \leq c_1|z|^\beta \psi(|z|); \tag{3.15}$$

(O) *For all $z \notin S_\theta$ with $|z| \geq z_0$,*

$$|\widetilde{G}(z)e^z| \leq c_2 e^{a|z|}. \tag{3.16}$$

Then for $n \geq 1$,

$$g_n = \widetilde{G}(n) + O\left(n^{\beta-1}\psi(n)\right) . \tag{3.17}$$

More precisely,

$$g_n = \widetilde{G}(n) - \tfrac{1}{2}n\widetilde{G}''(n) + O\left(n^{\beta-2}\psi(n)\right) . \tag{3.18}$$

The "Big-Oh" terms in (3.17) and (3.18) are uniform for any family of entire functions $\widetilde{G}$ that satisfy the conditions with the same a, β, θ, c_1, c_2, z_0 and ψ.

Now, we are in a position to depoissonize our results. By (3.13), it follows that

$$\begin{aligned} |\widetilde{W}_j(z)| &= \left|\Gamma(j)\frac{q(p/q)^j + p(q/p)^j}{h} \right.\\ &\qquad \left. + \delta_j(\log_{1/p} z) + O(z^{-M})\right| \\ &\leq \left|\Gamma(j)\frac{q(p/q)^j + p(q/p)^j}{h}\right| + \left|\delta_j(\log_{1/p} z)\right| \\ &\qquad + O(|z|^{-M}) \\ &= O(1) \end{aligned}$$

since $|\delta_j|$ is uniformly bounded on $\mathbb{C}$.

By (3.14), we see that

$$\begin{aligned} |\widetilde{G}(z,u)| &= \left| -\frac{q\ln(1-pu) + p\ln(1-qu)}{h} \right.\\ &\qquad \left. + \delta(\log_{1/p} z, u) + O(z^{-M})\right| \\ &\leq \left|-\frac{q\ln(1-pu) + p\ln(1-qu)}{h}\right| \\ &\qquad + \left|\delta(\log_{1/p} z, u)\right| + O(|z|^{-M}) \\ &= O(1) \end{aligned}$$

when u is fixed since $|\delta|$ is uniformly bounded on $\mathbb{C}$.

We define $\psi(z) = 1\ \forall z$ and note ψ is a slowing varying function (i.e., $\psi : (0,\infty) \to (0,\infty)$ and $\psi(tx)/\psi(x) \to 1$ as $x \to \infty$ for every fixed $t > 0$). Also there exist real-valued constants c_M, $c_{j,M}$, z_M, $z_{j,M}$ such that

$$|\widetilde{W}_j(z)| \leq c_{j,M}|z|^0\psi(|z|)$$

$\forall z \in S_{\pi/4} = \{z : |\arg(z)| \leq \pi/4\}$ with $|z| \geq z_{j,M}$, and

$$|\widetilde{G}(z,u)| \leq c_M|z|^0\psi(|z|)$$

$\forall z \in S_{\pi/4} = \{z : |\arg(z)| \leq \pi/4\}$ with $|z| \geq z_M$. So condition (I) of Theorem 3.1 is satisfied. It follows immediately by Theorem 10.4 of [17] that condition (O) of [17] (see page 456) is also satisfied. So by Theorem 3.1 it follows that Theorems 1 and 2 hold, as claimed.

To see that (2.5) follows from (2.4), consider the following. From (2.4), we have $E[u^{M_n}] =$

$$\begin{aligned} E[u^{M_n}] &= -\frac{q\ln(1-pu) + p\ln(1-qu)}{h} \\ &= \delta(\log_{1/p} n, u) + O(n^{-1}). \end{aligned} \tag{3.19}$$

Observe

$$-\frac{q\ln(1-pu) + p\ln(1-qu)}{h} = \sum_{j=1}^{\infty}\left(\frac{p^j q + q^j p}{jh}\right)u^j.$$

Also note that

$$\begin{aligned} &\delta(\log_{1/p} n, u) \\ &\quad = \sum_{k\neq 0} -\frac{e^{2kr\pi i \log_{1/p} n}\Gamma(z_k)}{p^{-z_k+1}\ln p + q^{-z_k+1}\ln q} \\ &\qquad\quad \times \left(q(1-pu)^{-z_k} + p(1-qu)^{-z_k} \right.\\ &\qquad\qquad \left. -p^{-z_k+1} - q^{-z_k+1}\right) \\ &\quad = \sum_{j=1}^{\infty}\sum_{k\neq 0} -\frac{e^{2kr\pi i \log_{1/p} n}\Gamma(z_k)(p^j q + q^j p)(z_k)^{\overline{j}}}{j!(p^{-z_k+1}\ln p + q^{-z_k+1}\ln q)}u^j. \end{aligned}$$

Then we apply these observations to (3.19) to conclude that (2.5) holds.

Finally, we note that (2.6) is an immediate corollary of (2.5).

References

[1] Lars V. Ahlfors, *Complex Analysis*, 3rd ed. (New York: McGraw-Hill, 1979).

[2] William Feller, *An Introduction to Probability Theory and Its Applications*, Volume I, 3rd ed., Volume II, 2nd ed. (New York: Wiley, 1968, 1971).

[3] James Allen Fill, Hosam M. Mahmoud, and Wojciech Szpankowski, "On the distribution for the duration of a randomized leader election algorithm," *The Annals of Applied Probability* **6** (1996), 1260–1283.

[4] Philippe Flajolet and Robert Sedgewick, *Analytic Combinatorics* (forthcoming). [A preview is available at <http://algo.inria.fr/flajolet/Publications/books.html>.]

[5] Philippe Flajolet, Xavier Gourdon, and Philippe Dumas, "Mellin transforms and asymptotics: Harmonic sums," *Theoretical Computer Science* **144** (1995), 3–58.

[6] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik, *Concrete Mathematics*, 2nd ed. (Reading, Massachusetts: Addison-Wesley, 1994).

[7] Daniel H. Greene and Donald E. Knuth, *Mathematics for the Analysis of Algorithms*, 3rd ed. (Boston: Birkhäuser, 1990).

[8] Peter Henrici, *Applied and Computational Complex Analysis*, Volumes 1–3. (New York: Wiley, 1974, 1977, 1986).

[9] Philippe Jacquet and Wojciech Szpankowski, "Autocorrelation on words and its applications. Analysis of suffix trees by string-ruler approach," *Journal of Combinatorial Theory* **A66** (1994), 237–269.

[10] Philippe Jacquet and Wojciech Szpankowski, "Analytical depoissonization and its applications," *Theoretical Computer Science* **201** (1998), 1–62.

[11] Carl Svante Janson and Wojciech Szpankowski, "Analysis of an asymmetric leader election algorithm," *The Electronic Journal of Combinatorics* **4(1)**, R17 (1997), 1–16.

[12] Donald E. Knuth, *Fundamental Algorithms*, 3rd ed. (Reading, Massachusetts: Addison Wesley, 1997).

[13] Stefano Lonardi and Wojciech Szpankowski, "Joint source-channel LZ'77 coding," *Data Compression Conference* (2003), 273–282.

[14] Helmut Prodinger, "How to select a loser," *Discrete Mathematics* **120** (1993), 149–159.

[15] Robert Sedgewick and Philippe Flajolet, *An Introduction to the Analysis of Algorithms* (Reading, Massachusetts: Addison-Wesley, 1996).

[16] Wojciech Szpankowski, "A generalized suffix tree and its (un)expected asymptotic behaviors," *SIAM Journal on Computing* **22** (1993), 1176–1198.

[17] Wojciech Szpankowski, *Average Case Analysis of Algorithms on Sequences* (New York: Wiley, 2001).

[18] Jacob Ziv and Abraham Lempel, "A universal algorithm for sequential data compression," *IEEE Transactions on Information Theory* **IT-23** (1977), 337–343.