ON THE ANALYSIS OF PROBABILISTIC COUNTING

PETER KIRSCHENHOFER AND HELMUT PRODINGER

Department of Algebra and Discrete Mathematics
Technical University of Vienna, Austria

Abstract. In this note an alternative analysis of a probabilistic counting algorithm due to Flajolet and Martin is presented. The asymptotic evaluation of certain combinatorial sums is performed via residue calculus instead of Flajolet's Mellin transform approach that had to use some unpleasant real analysis.

1. INTRODUCTION

A basic probabilistic counting procedure to estimate the number n of distinct elements in a multiset M uses the following idea (compare [**1**]): We map the possible domain into the set of sufficiently long strings of 0 and 1 such that each string is taken as a value of this mapping with equal probability (a so-called "hash-function"). The bitwise OR-composition of all the images of the elements of M contains information on the size of n: the position of the leftmost zero is used as an estimate of $\log_2 n$. Let R be this quantity. In [**1**] it is shown that the expected value $\overline{R_n}$ of R fulfills

THEOREM 1.

$$\overline{R_n} \sim \log_2 n + C_1 + \delta_1(\log_2 n)$$

$$\textit{where } C_1 = -\frac{1}{2} + \frac{\gamma}{\log 2} + \log\left(\frac{2}{3}\prod_{p=1}^{\infty}\left[\frac{(4p+1)(4p+2)}{4p(4p+3)}\right]^{(-1)^{\nu(p)}}\right),$$

$$\nu(p) = \textit{the number of ones in the binary representation of } p,$$

$$\textit{and } \delta_1(x) = \frac{1}{\log 2}\sum_{k\neq 0}\Gamma\left(\frac{2k\pi i}{\log 2}\right) N\left(\frac{2k\pi i}{\log 2}\right),$$

$$\textit{with } N(s) = \textit{ the analytic continuation of } \sum_{j\geq 1}\frac{(-1)^{\nu(j)}}{j^s}.$$

We mention that $\delta_1(x)$ is a continuous periodic function of period 1 with very small amplitude and mean zero, so that for practical purposes this periodic fluctuations may be safely ignored.

The variance σ_n^2 of the random variable R is very small:

THEOREM 2.

$$\sigma_n^2 \sim \frac{\pi^2}{6\log^2 2} - \frac{N'(0)^2}{\log^2 2} - \frac{N''(0)}{\log^2 2} + \frac{1}{12} - [\delta_1^2]_0 + \delta_2(\log_2 n),$$

where $N(s)$ is defined in Theorem 1, $[\delta_1^2]_0$ is the mean of $\delta_1^2(x)$, and $\delta_2(x)$ is again a periodic function of very small amplitude and mean zero.

The second author wants to thank M. Drmota for his advice and patience concerning TEX.

Numerically

$$\sigma_n^2 \sim 1.257\ldots\ .$$

Observe the somewhat erroneous expression for σ_n^2 in [**1**] whereas the numerical approximation therein is correct.

In [**1**] the Theorems are achieved by evaluating certain combinatorial sums by the use of Mellin's integral transform.

In this note we gain the results by a more concise method which has proved to be successful already in a number of other situations ([**2**],[**3**],[**4**]).

2. The Analysis

Let $q_{n,k}$ be the probability that for a multiset of n distinct elements $R \geq k$. In [**1**] it is shown that

$$q_{n,k} = \sum_{j=0}^{2^k-1} (-1)^{\nu(j)} \left(1 - \frac{j}{2^k}\right)^n . \tag{1}$$

Observing that

$$\overline{R_n} = \sum_{l \geq 1} q_{n,l} \tag{2}$$

and

$$\sigma_n^2 = 2 \sum_{l \geq 1} l\, q_{n,l} - \overline{R_n} - \overline{R_n}^2 , \tag{3}$$

we have to evaluate certain alternating sums asymptotically.

For $\overline{R_n}$ we have from (2) and (1)

$$\overline{R_n} = \sum_{k=1}^{n} \binom{n}{k} (-1)^k \sum_{l \geq 1} \sum_{j=1}^{2^l-1} (-1)^{\nu(j)} \left(\frac{j}{2^l}\right)^k ,$$

so that $\overline{R_n}$ is of the form

$$\overline{R_n} = \sum_{k=1}^{n} \binom{n}{k} (-1)^k f(k), \tag{4}$$

with

$$f(z) = \sum_{l \geq 1} \sum_{j=1}^{2^l-1} (-1)^{\nu(j)} \left(\frac{j}{2^l}\right)^z .$$

Using residue calculus it can be seen that the sum in (4) may be rewritten as a contour integral:

$$\overline{R_n} = -\frac{1}{2\pi i} \int_C [n;z] f(z) dz, \tag{5}$$

$$\text{with} \quad [n;z] = \frac{(-1)^{n-1} n!}{z(z-1)\cdots(z-n)},$$

where C surrounds the singularities $z = 1, 2, \ldots, n$ of the integrand.

Expanding the contour of integration in a well-suited manner (compare [4] for technical details), we find

$$\overline{R_n} \sim \sum \mathrm{Res}_{z=z_i}\Big([n;z]f(z)\Big) \tag{6}$$

where the sum is taken over all poles z_i of $[n;z]f(z)$ with real part larger than some fixed c that differ from $1,...,n$.

In our instance we may write

$$f(z) = f_1(z) - f_2(z),$$

where

$$\begin{aligned} f_1(z) &= \sum_{l\geq 1} \lim_{k\to\infty} \sum_{j=1}^{2^k-1} (-1)^{\nu(j)} \left(\frac{j}{2^l}\right)^z \\ &= \frac{1}{2^z-1} N(-z), \end{aligned}$$

and

$$f_2(z) = \sum_{l\geq 1} \lim_{k\to\infty} \sum_{j=2^l}^{2^k-1} (-1)^{\nu(j)} \left(\frac{j}{2^l}\right)^z .$$

From [1] it follows immediately that $f_2(z)$ is analytic for $\Re z > -1$ with $f_2(0) = 0$.

If we use $c = -\frac{1}{2}$ in (6) and take into account the residues at the poles $z = \chi_k = 2k\pi i/L$, $k \in \mathbf{Z}$; $L = \log 2$, of $[n;z]f_1(z)$ we find:

i) For $z \to 0$

$$\begin{aligned} &\frac{1}{2^z-1} \sim \frac{1}{Lz} - \frac{1}{2}, \\ &N(-z) \sim N(0) - zN'(0), \quad \text{where} \quad N(0) = -1, \\ &[n;z] \sim -\frac{1}{z} - H_n, \quad \text{where} \quad H_n = \sum_{k=1}^{n} \frac{1}{k}, \end{aligned}$$

so that

$$\mathrm{Res}_{z=0}\Big([n;z]f_1(z)\Big) \sim \frac{\log n}{L} + \left(\frac{\gamma}{L} + \frac{N'(0)}{L} - \frac{1}{2}\right)$$

ii) For $z \to \chi_k, \quad k \neq 0$

$$\begin{aligned} \frac{1}{2^z-1} &\sim \frac{1}{L(z-\chi_k)}, \\ N(-z) &\sim N(-\chi_k), \\ [n;z] &\sim n^{\chi_k}\Gamma(-\chi_k), \end{aligned}$$

so that

$$\mathrm{Res}_{z=\chi_k}\Big([n;z]f_1(z)\Big) \sim \frac{1}{L} N(-\chi_k)\Gamma(-\chi_k)e^{2k\pi i \log_2 n}.$$

iii) Since $f_2(0) = 0$, we have

$$\mathrm{Res}_{z=0}\Big([n;z]f_2(z)\Big) = 0.$$

Combining these results we rederive Theorem 1.

For the evaluation of $N'(0)$ we use the following formula

$$N'(0) = \log\frac{2}{3} + \sum_{p\geq 1}(-1)^{\nu(p)}\log\frac{(4p+1)(4p+2)}{4p(4p+3)} \tag{7}$$

which occurs by grouping terms four by four.

The analysis of the variance follows the same idea. We have

$$\sum_{l\geq 1} l\, q_{n,l} = \sum_{k=1}^{n}\binom{n}{k}(-1)^k g(k)$$

with

$$g(z) = \frac{2^z}{(2^z-1)^2}N(-z) - g_2(z),$$

where $g_2(z)$ is analytic for $\Re z > -1$ and fulfills $g_2(0) = 0$.

Calculating the residues of

$$[n;z]\frac{2^z}{(2^z-1)^2}N(-z)$$

we find that, apart from periodic fluctuations of mean zero,

$$\begin{aligned}\sum_{l\geq 1} l\, q_{n,l} \sim \frac{1}{2}\frac{\log^2 n}{L^2} + \left\{\frac{\gamma}{L^2} + \frac{N'(0)}{L^2}\right\}\log n \\ + \frac{\gamma^2}{2L^2} + \frac{\pi^2}{12L^2} + \frac{\gamma N'(0)}{L^2} - \frac{N''(0)}{2L^2} - \frac{1}{12}.\end{aligned} \tag{8}$$

From (8) we gain Theorem 2 immediately.

References

1. P. Flajolet and G.N. Martin, *Probabilistic Counting Algorithms for Data Base Applications*, J.Comput.Syst.Sci. **31** (1985), 182–209.
2. P. Flajolet and R. Sedgewick, *Digital Search Trees Revisited*, SIAM J.Comput. **15** (1986), 748–767.
3. P. Kirschenhofer and H. Prodinger, *Approximate Counting: An Alternative Analysis*, RAIRO Informatique Théorique (1990) (to appear).
4. U. Schmid, *Analyse von Collision-Resolution Algorithmen in Random-Access- Systemen mit dominanten Übertragungskanälen*, Dissertation TU Wien (1986).

TU Vienna, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria