NOTE

HYPOTHETIC ANALYSES: APPROXIMATE COUNTING IN THE STYLE OF KNUTH, PATH LENGTH IN THE STYLE OF FLAJOLET

Helmut Prodinger

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

October 16, 1991

1. Introduction

The first analysis of *approximate counting* is due to Flajolet [1], whereas the first satisfactory analysis of the *average path length in digital search trees* has been performed by Knuth [4,p.497ff]. Both authors have used the *Mellin Integral Transform* (see [2] for a brilliant introduction), but in rather different ways. It was shown by Kirschenhofer and the present writer [3] that both problems are very similar (not to say "almost identical").

It is now amusing to figure out what Flajolet and Knuth would have done by considering the exchanged problems. The aim of this note is to perform their analyses. It is the author's opinion that this is entertaining and amusing both from the methodological as well as the *historical* point of view.

2. Flajolet's analysis of approximate counting

Flajolet considered $C_N^{[F]}$, the average value of a counter after N increments in a probabilistic algorithm. He computed $C_N^{[F]}$ via

$$C_N^{[F]} = \sum_{l\ge 1} l \cdot p_{N,l}^{[F]} \ ,$$

with

$$p_{N,l}^{[F]} = \sum_{j=0}^{l-1} \frac{(-1)^j 2^{-\binom{j}{2}} \left(1-2^{j-l}\right)^N}{Q_j Q_{l-1-j}}$$

and

$$Q_m = \prod_{i=1}^{m} \left(1-\frac{1}{2^i}\right) .$$

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

Then he approximated $p^{[F]}_{N,l}$ by $\Phi\left(\frac{N}{2^l}\right)$, with

$$\Phi(x) = \frac{1}{Q_\infty}\sum_{j\ge 0}\frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j}e^{-x2^j}.$$

Here,

$$Q_\infty = \lim_{m\to\infty} Q_m = \prod_{i\ge 1}\left(1-\frac{1}{2^i}\right) = 0.288788\cdots.$$

Then the Mellin transform was computed:

$$\Phi^*(s) = \Gamma(s)\cdot\xi(s),$$

with

$$\xi(s) = \frac{1}{Q_\infty}\sum_{j\ge 0}\frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j}2^{-js}.$$

Finally,

$$C^{[F]}_N \sim F(N) = \sum_{l\ge 1} l\,\Phi\left(\frac{N}{2^l}\right),$$

and

$$F^*(s) = \frac{2^s}{(2^s-1)^2}\Phi^*(s).$$

By considering the *negative residues* of $F^*(s)x^{-s}$ for $\Re s = 0$ he obtained

Theorem 1. *(Flajolet, 1985* [1]*)*

$$C^{[F]}_N \sim \log_2 N + \frac{\gamma}{\log 2} - \alpha + \frac{1}{2} - \delta_0\left(\log_2 N\right),$$

where

$$\begin{aligned}
\alpha &= \sum_{i\ge 1}\frac{1}{2^i-1} = 1.6066\cdots, \\
\gamma &= 0.57721\cdots \qquad \textit{(Euler's constant)}, \\
\delta_0(x) &= \frac{1}{\log 2}\sum_{k\ne 0}\Gamma\left(-\frac{2k\pi i}{\log 2}\right)e^{2k\pi i x}. \quad \Box
\end{aligned}$$

3. Knuth's analysis of the average path length in DSTs

Knuth considered

$$L^{[K]}_N = \sum_{k=2}^{N}\binom{N}{k}(-1)^k Q_{k-2}. \tag{K}$$

He changed it to

$$L^{[K]}_N \sim \sum_{n\ge 0} a_{n+1}\sum_{m\ge 0}2^{m(1-n)}\left(e^{-N/2^m} - 1 + \frac{N}{2^m}\right),$$

with

$$a_i = \frac{(-1)^{i+1} 2^{-\binom{i}{2}}}{Q_{i-1}}.$$

By "Mellin"[1] he obtained

$$L_N^{[K]} \sim \sum_{n \geq 0} a_{n+1} \frac{1}{2\pi i} \int_{-\frac{3}{2}-i\infty}^{-\frac{3}{2}+i\infty} \Gamma(s) N^{-s} \frac{1}{1-2^{1-n+s}} ds.$$

The negative residues of the integrand for $\Re s = -1$ give the following result

Theorem 2. *(Knuth, 1973* [4]*)*

$$L_N^{[K]} \sim N \left(\log_2 N + \frac{\gamma}{\log 2} - \alpha + \frac{1}{2} - \frac{1}{\log 2} + \delta_{-1}(\log_2 N) \right),$$

with

$$\delta_{-1}(x) = \frac{1}{\log 2} \sum_{k \neq 0} \Gamma\left(-1 - \frac{2k\pi i}{\log 2}\right) e^{2k\pi i x}. \quad \square$$

Now we could show [3] that

$$C_N^{[F]} = 1 - \sum_{k=1}^{N} \binom{N}{k} (-1)^k 2^{-k} Q_{k-1}, \tag{F}$$

and this similarity with (K) motivated this note. Section 6 contains an attempt to "explain" this phenomen.

4. Knuth's analysis of approximate counting

We need some "Q-formulæ" (see, e.g. [1], [4, exercise 5.1.1–16]). Let

$$Q(x) = \prod_{i \geq 1} \left(1 - \frac{x}{2^i}\right),$$

then

$$Q_k = \frac{Q_\infty}{Q(2^{-k})}. \tag{Q1}$$

Euler's formulæ:

$$\frac{1}{1-t} \frac{1}{Q(t)} = \sum_{n \geq 0} \frac{t^n}{Q_n} \tag{E1}$$

---

[1] Actually, Knuth never mentioned the word "Mellin transform". He basically used the formula $e^{-x} = \frac{1}{2\pi i} \int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} \Gamma(s) x^{-s} ds$ and some variants that can be proved directly by residue calculus. Because of the *ubiquitous* appearance of the Gamma function he christened the approach "Gamma function method". Also, he gives the credits for this procedure to N. G. de Bruijn.

$$Q(t)=\sum_{n\ge0}a_{n+1}t^n \tag{E2}$$

Finally we need

$$\sum_{l\ge1}\frac{a_{l+1}}{1-2^{-l}}=-\alpha;$$

this follows from

$$\frac{Q_\infty}{Q(x)}=\sum_{i\ge1}\frac{a_i}{1-x2^{-i}}.$$

Now

$$\begin{aligned}
C_N^{[F]}-1&=-Q_\infty\sum_{k=1}^N\binom{N}{k}(-1)^k2^{-k}\prod_{l\ge0}\frac{1}{1-2^{-l-k}} && \text{(by Q1)}\\
&=-Q_\infty\sum_{k=1}^N\binom{N}{k}(-1)^k2^{-k}\sum_{m\ge0}2^{-km}\frac{1}{Q_m} && \text{(by E1)}\\
&=-\sum_{m\ge0}\left(\sum_{k=0}^N\binom{N}{k}(-1)^k2^{-k(m+1)}-1\right)Q\left(2^{-m}\right) && \text{(by Q1)}\\
&=-\sum_{m\ge0}\left(\left(1-2^{-m-1}\right)^N-1\right)\sum_{n\ge0}a_{n+1}2^{-mn} && \text{(by E2)}\\
&\sim-\sum_{n\ge0}a_{n+1}\sum_{m\ge0}2^{-mn}\left(e^{-N/2^{m+1}}-1\right)\\
&=-\sum_{n\ge0}a_{n+1}\frac{1}{2\pi i}\int_{-\frac12-i\infty}^{-\frac12+i\infty}\Gamma(s)N^{-s}\frac{2^s}{1-2^{-n+s}}ds.
\end{aligned}$$

The negative residues at $s=0$ and $s=\frac{2k\pi i}{\log 2}$ $(k\ne0)$ give Flajolet's result: From $n=0$ we obtain

$$\log_2 N+\frac{\gamma}{\log 2}-\frac12-\delta_0\left(\log_2 N\right);$$

from $n\ge1$ we obtain

$$\sum_{n\ge1}a_{n+1}\frac{1}{1-2^{-n}}=-\alpha.$$

5. Flajolet's analysis of the average path length in DSTs

We want to go back from (K) to

$$L_N^{[K]}=\sum_{l\ge1}l\cdot p_{N,l}^{[K]}\ ;$$

this representation is of course not unique. Let

$$g(x)=\prod_{m=0}^{k-2}\left(1-\frac{x}{2^m}\right).$$

Then

$$g'(1)=\begin{cases}0 & \text{for} \quad k=0\\ -Q_{k-2} & \text{for} \quad k\geq 1.\end{cases}$$

Furthermore, we have in general

$$\sum_{l\geq 1} l\left[x^l\right] g(x)=g'(1),$$

where $\left[x^l\right] g(x)$ denotes the coefficient of x^l in the Taylor series expansion of $g(x)$.
Thus

$$L_N^{[K]}=-\sum_{k=1}^{N}\binom{N}{k}(-1)^k\sum_{l\geq 1} l\left[x^l\right]\prod_{m=0}^{k-2}\left(1-\frac{x}{2^m}\right).$$

By (E1) and (E2) we have

$$\begin{aligned}\prod_{m=0}^{k-2}\left(1-\frac{x}{2^m}\right)&=\prod_{m\geq 0}\left(1-\frac{x}{2^m}\right)\prod_{m\geq 0}\left(1-\frac{x}{2^{m+k-1}}\right)^{-1}\\ &=\sum_{j\geq 0}\frac{(-1)^j 2^{-\binom{j}{2}}x^j}{Q_j}\sum_{s\geq 0}\frac{1}{Q_s}\left(2^{1-k}\right)^s x^s\end{aligned}$$

and thus

$$\left[x^l\right]\prod_{m=0}^{k-2}\left(1-\frac{x}{2^m}\right)=\sum_{j=0}^{l}\frac{(-1)^j 2^{-\binom{j}{2}}}{Q_jQ_{l-j}}2^{-(l-j)(k-1)}.$$

So we can write

$$L_N^{[K]}=\sum_{l\geq 1} l\cdot p_{N,l}^{[K]},$$

with

$$\begin{aligned}p_{N,l}^{[K]}&=-\sum_{k=1}^{N}\binom{N}{k}(-1)^k\sum_{j=0}^{l}\frac{(-1)^j 2^{-\binom{j}{2}}}{Q_jQ_{l-j}}2^{-(l-j)(k-1)}\\ &=\sum_{j=0}^{l}\frac{(-1)^j 2^{-\binom{j}{2}}}{Q_jQ_{l-j}}2^{l-j}\left[1-\left(1-2^{-(l-j)}\right)^N\right].\end{aligned}$$

Flajolet would approximate $p_{N,l}^{[K]}/N$ by $\Psi\left(\frac{N}{2^l}\right)$, with

$$\Psi(x)=\frac{1}{Q_\infty}\sum_{j\geq 0}\frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j}\frac{1}{x2^j}\left(1-e^{-x2^j}\right).$$

Now we perform the Mellin transform and find for $-1<\Re s<0$

$$\left(\frac{1-e^{-x}}{x}\right)^*(s)=\int_0^\infty\left(1-e^{-x}\right)x^{s-2}dx=-\Gamma(s-1).$$

Therefore

$$\Psi^*(s) = -\xi(s)\Gamma(s-1),$$

where the function $\xi(s)$ was defined in Section 2. Now

$$\frac{L_N^{[K]}}{N} \sim K(N)$$

with

$$K(x) = \sum_{l \ge 1} l\,\Psi\left(\frac{x}{2^l}\right).$$

As before,

$$K^*(s) = \frac{2^s}{(2^s-1)^2} \cdot \Psi^*(s).$$

By Mellin's inversion theorem we have

$$K(x) = -\frac{1}{2\pi i} \int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \frac{2^s}{(2^s-1)^2} \cdot \xi(s) \cdot \Gamma(s-1) \cdot x^{-s}\, ds.$$

The negative residue of the integrand at $s = 0$ gives the contribution

$$\log_2 N - \frac{1}{\log 2} + \frac{\gamma}{\log 2} - \alpha + \frac{1}{2};$$

the residues at $s = \frac{2k\pi i}{\log 2}$ produce the contribution $\delta_{-1}(\log_2 N)$.

6. An "explanation" of the similarities of the two problems

After a first reading of this paper (1991), an anonymous referee asked for a combinatorial explanation of the surprising similarity of the two seemingly different problems. This is of course a natural question and I failed in 1990 to give a good answer. However now I can offer something.

The *probability generating functions* for the path length in digital search trees obey the following recursion ($N \ge 1$, $F_1(z) = F_0(z) = 1$)

$$F_{N+1}(z) = z^N \sum_{k=0}^{N} 2^{-N} \binom{N}{k} F_k(z) F_{N-k}(z).$$

(The averages L_N studied by Knuth are obtained via $L_N = F_N'(1)$.)

Now I will give a combinatorial exlanation of a similar recursion for the approximate counting problem. Let, as in Section 1, $p_{N,l}^{[F]}$ denote the probability that the counter has the value l ("level l") after N "inputs", and

$$F_N(z) = \sum_{l \ge 0} p_{N,l}^{[F]}\, z^l.$$

It should be noted that the counter is initially set to 1. To have a good idea what's going on, let's say that there are N "players"; whenever he (or she) has his turn, he tries to increase the counter by 1. If the counter has the value l, he will succeed with probability 2^{-l}; otherwise his attempt has no effect. For instance, we can ask him to throw l consecutive "heads" (or 1's) with a fair coin. But we can think about it in another way: He starts at level 1 and flips a fair coin and can proceed if he has a 1. If he reaches the level $l+1$, it's exactly what we need. Now we no longer have to think about N *consecutive* players; each of them can throw the dice as long as he has 1's; throwing a 0 means to stop. How is the value of counter computed from these data? We start at level 1 and, being at level j, we can go on if there are at least j players who made it to level $j+1$. The maximal level that we reach in that way is the value of the counter. Indeed, to proceed to a higher level means to kick out a successful player. From this interpretation we can set up a recursion for the probability generating functions. If $k \ge 1$ players have been successful at level 1, $k-1$ are going to continue. But if $k=0$, the recursion stops. Hence

$$F_N(z) = z\sum_{k=1}^{N} 2^{-N}\binom{N}{k}F_{k-1}(z) + z2^{-N}, \quad N \ge 1,\ F_0(z) = z.$$

Now set, as usual, $C_N = F_N'(1)$. Then the recursion turns into

$$C_N = 1 + 2^{-N}\sum_{k=1}^{N}\binom{N}{k}C_{k-1}, \quad N \ge 1,\ C_0 = 1.$$

If we subtract the analogous recursion with N replaced by $N-1$, it turns into the more feasible

$$2C_N = C_{N-1} + 1 + 2^{1-N}\sum_{k=0}^{N-1}\binom{N-1}{k}C_k, \quad N \ge 1,\ C_0 = 1.$$

The standard method to solve this recursion is by the use of the exponential generating function $C(z) = \sum_{N\ge 0} C_N z^N/N!$. Then

$$2C'(z) = C(z) + e^z + e^{z/2}C\left(\frac{z}{2}\right).$$

Now one sets $D(z) = e^{-z}C(z)$ to obtain the easier equation

$$2D'(z) = -D(z) + D\left(\frac{z}{2}\right) + 1.$$

Then we consider the coefficients:

$$2D_{N+1} = -\left(1-2^{-N}\right)D_N, \quad N \ge 1,\ D_0 = 1,\ D_1 = \frac{1}{2}.$$

Iterating this we find

$$D_N = 2^{-N}(-1)^{N-1}Q_{N-1}, \quad N \ge 1,\ D_0 = 1.$$

Since $C(z) = e^z D(z)$, we have

$$C_N = \sum_{k=0}^{N}\binom{N}{k}D_k = 1 - \sum_{k=1}^{N}\binom{N}{k}(-1)^k 2^{-k}Q_{k-1};$$

this is formula (F).

Similar ideas are used in a forthcoming paper on probabilistic counting algorithms by Peter Kirschenhofer, Wojciech Szpankowski and the author.

7. Conclusion

Now it is worthwile to discuss the two "Mellin" styles. Knuth and Flajolet attacked their problems from different sides (both of them very natural in their respective contexts). Since in "approximate counting" it is natural to compute the probabilities directly, it was a good idea to approximate them immediately and use "Mellin" after that. In the tree context the recursions (as in Section 6) are most natural, so that Knuth *had* to use Euler's formulæ in order to avoid the cancellations in (K). After that "Mellin" can be used. Flajolet used first "Mellin" and then "Euler" (in order to analyze the ξ–function); in Knuth's case it is the other way around.

It should be stated here explicitly that this note contains no *judgement* about the relative qualities of the approaches. Both of them are most valuable and have influenced many people.

References

1. P.Flajolet, *Approximate Counting: A Detailed Analysis*, BIT **25** (1985), 113–134.
2. P.Flajolet and J.Vitter, *Average-Case Analysis of Algorithms and Data Structures*, Handbook of Theoretical Computer Science Vol. A "Algorithms and Complexity", North-Holland, 1990, pp. 431–524.
3. P.Kirschenhofer and H. Prodinger, *Approximate Counting: An Alternative Approach*, RAIRO Theoretical Informatics **25** (1991), 43–48.
4. D.E.Knuth, *The Art of Computer Programming Vol. 3*, Addison-Wesley, Reading MA, 1973.

TU Vienna
Wiedner Hauptstrasse 8-10
A-1040 Vienna
Austria