Some Analytic Techniques for the Investigation of the Asymptotic Behaviour of Tree Parameters

Helmut Prodinger

Department of Algebra and Discrete Mathematics
Technical University of Vienna
Wiedner Hauptstraße 8–10
A–1040, Vienna, Austria

Written for the Proceedings of the Summerschool in Ile de Ré 1983 (in French)
Retyped 1991
Translated into English January 1992

1. Introduction

The structure of a tree is one of the most common data structures in Computer Science. There exist many algorithms for the manipulation of trees. The analysis of these algorithms in the sense of Knuth leads often to combinatorial problems of the enumeration of configurations; the following step consists of the asymptotic analysis of these numbers. See e.g. [4], [15], [16].

The study of Flajolet and Odlyzko [7] is a paradigm for a slightly different point of view: Instead of extracting the coefficients of generating functions and approximating them by methods of real analysis, the singularities of the generating functions are studied. More precisely, one tries to establish the local behaviour of the generating functions around their singularities next to the origin. After that, some translation lemmas allow to translate this behaviour into the asymptotic behaviour of the (Taylor-)coefficients of the generating function in question. For instance, if $f(z)=\sum_{n\ge0} f_n z^n$ has its singularity next to the origin at 1, with

$$f(z)\sim A-B\sqrt{1-z}+\dots,$$

then one concludes that

$$[z^n]\,f(z):=f_n\sim\frac{B}{2\sqrt{\pi}}n^{-3/2}.$$

Similarily if

$$f(z)\sim -A\log(1-z)+\dots,$$

then

$$[z^n]f(z)\sim\frac{A}{n},\quad \text{etc.}$$

Also, the more terms we know from the local behaviour of the generating function, the more terms we know from the asymptotic behaviour of the (Taylor-)coefficients.

This translation is a rather automatic process. The principle problem is thus to find the local behaviour around the (dominant) singularities. In a lot of cases one can use the Mellin transform, a very popular method within the french school of the analysis of algorithms.

The main interest of this paper is to explain this general strategy in 3 particular cases: the problem of register allocation for the evaluation of arithmetic expressions; then the height of leaves in planted plane trees. The third problem is the asymptotic study of a certain expression (that is similar to one occurring in the register problem); this value is of interest in the analysis of a particular sorting algorithm (see [6], [21]).

Before we engage ourselves into technical problems, we would like to formulate some fundamental concepts:

The family $\mathcal{B}$ of binary trees might be defined recursively by the (symbolic) equation

which means that a binary tree is either an external node ("leaf") or built by a root and 2 binary trees (the left and right subtree). Let $B(z) = \sum_{n\geq 0} b_n z^n$ denote the generating function where b_n is the number of binary trees of size n (the size is measured in termes of internal nodes); one can see immediately that

$$B(z) = 1 + z\,(B(z))^2\,,$$

and thus

$$B(z) = \frac{1-\sqrt{1-4z}}{2z} \quad \text{and} \quad b_n = \frac{1}{n+1}\binom{2n}{n}.$$

The numbers b_n are *Catalan* numbers.

In an analogous way, the family $\mathcal{P}$ of planted plane trees is defined by

form which one deduces the generating function $P(z) = \sum_{n\geq 1} t_n z^n$ of planted plane trees:

$$P(z) = \frac{z}{1-P(z)} \quad \text{and thus} \quad P(z) = \frac{1-\sqrt{1-4z}}{2} \quad \text{and} \quad t_n = \frac{1}{n}\binom{2n-2}{n-1}.$$

In these 2 cases, the above mentioned singularity is $z = \frac{1}{4}$, and

$$B(z) \sim 2 - 2\sqrt{1-4z}, \quad \text{thus} \quad b_n \sim \frac{1}{\sqrt{\pi}} 4^n n^{-3/2}$$

or

$$P(z) \sim \frac{1}{2} - \frac{1}{2}\sqrt{1-4z}, \quad \text{thus} \quad t_n \sim \frac{1}{\sqrt{\pi}} 4^{n-1} n^{-3/2}.$$

2. The average number of registers to evaluate a binary tree of size n

Before we start the discussion, let's cite the references on this subject: [4], [5], [8], [11], [17], [19]. We won't go into a description of the Computer Science problem. For this we advise the gentle reader to consult the thesis of P. Flajolet [4].

The function "number of registers" *reg* is given recursively by

$$\operatorname{reg}(\square) = 0$$

$$\operatorname{reg}\left(\quad\right) = \begin{cases} \max\{\operatorname{reg}(t_1), \operatorname{reg}(t_2)\} & \text{if } \operatorname{reg}(t_1) \neq \operatorname{reg}(t_2) \\ 1 + \operatorname{reg}(t_1) & \text{otherwise.} \end{cases}$$

The problem is the evaluation of R_n, the average value of the function *reg*, based on the set of binary trees of size n. We have

$$R_n = \frac{1}{b_n} \cdot [z^n] E(z),$$

with

$$E(z) = \sum_{n \geq 0} z^n \sum_{t \text{ of size } n} \operatorname{reg}(t).$$

Let $\mathcal{R}_p$ be the family of trees t with $\operatorname{reg}(t) = p$. From the definition of *reg* we easily see that

Let $\mathcal{S}_p$ be the family of trees t with $\operatorname{reg}(t) \geq p$, that means

$$\mathcal{S}_p = \mathcal{R}_p + \mathcal{R}_{p+1} + \dots \ .$$

We can write

$$\sum_{j < p-1} \mathcal{R}_j = \mathcal{B} \setminus \mathcal{S}_{p-1}.$$

We obtain a recursion for the families $\mathcal{S}_p$ in a very similar fashion:

The functions $S_p(z)$ are defined by

$$S_p(z) := \sum_{n \geq 0} z^n \sum_{t \text{ of size } n, \ \operatorname{reg}(t) \geq p} 1.$$

It is clear that

$$E(z) = \sum_{p \geq 1} S_p(z).$$

The functions $S_p(z)$ may be described by the immediate recursions

$$S_0 = B; \quad S_p = z S_{p-1}^2 + 2z S_p \left(B - S_{p-1}\right).$$

It is convenient to define $\varepsilon := \sqrt{1-4z}$ and $z = u/(1+u)^2$. Thus

$$S_p = \frac{zS_{p-1}^2}{1-2zB+2zS_{p-1}},$$

or

$$\frac{1}{S_p} = \frac{\varepsilon}{zS_{p-1}^2} + \frac{2}{S_{p-1}}.$$

This finally gives us

$$\begin{aligned} 1+\frac{\varepsilon}{z}\frac{1}{S_p} &= 1+\frac{\varepsilon^2}{z^2S_{p-1}^2} + \frac{2\varepsilon}{z}\frac{1}{S_{p-1}} = \left(1+\frac{\varepsilon}{z}\cdot\frac{1}{S_{p-1}}\right)^2 \\ &= \cdots = \left(1+\frac{\varepsilon}{z}\cdot\frac{1}{S_0}\right)^{2^p}. \end{aligned}$$

From this we derive the equation

$$S_p = \frac{\varepsilon}{z}\left(\left(1+\frac{\varepsilon}{zB}\right)^{2^p} - 1\right)^{-1}.$$

Now we see at once that

$$\frac{\varepsilon}{z} = \frac{1-u^2}{u}, \quad \text{and} \quad 1+\frac{\varepsilon}{zB} = \frac{1}{u},$$

so that

$$S_p = \frac{1-u^2}{u}\frac{1}{u^{-2^p}-1} = \frac{1-u^2}{u}\frac{u^{2^p}}{1-u^{2^p}}.$$

For the function $E(z)$ we have therefore the following expression

$$\begin{aligned} E(z) &= \frac{1-u^2}{u}\sum_{p\geq 1}\frac{u^{2^p}}{1-u^{2^p}} \\ &= \frac{1-u^2}{u}\sum_{p\geq 1}\sum_{\lambda\geq 1} u^{\lambda 2^p} \\ &= \frac{1-u^2}{u}\sum_{n\geq 2} v_2(n)u^n, \end{aligned}$$

where $v_2(n)$ is the exponent of 2 in the prime number decomposition of n:

$$v_2(n) = \sum_{n=\lambda 2^p,\ \lambda,p\geq 1} 1.$$

If we write in a unique way $n = 2^m(1+2i)$, $(m\geq 1,\ i\geq 0)$, we have

$$v_2(n) = \sum_{p=1}^{m} 1 = m.$$

The quantity $v_2(n)$ is often called "dyadic valuation", for obvious reasons. As it was announced in the Introduction, we will finally show that in a neighbourhood of $z=\frac14$ (which means $u=1$):

$$E(z) \sim 2\varepsilon\log_2\varepsilon + A\varepsilon + 4\sum_{k\neq 0} c_k \varepsilon^{1-\chi_k} + \ldots,$$

with

$$A = 3 - 2\log_2 2\pi + \frac{2\gamma}{\log 2}, \quad (\gamma \text{ is Euler's constant})$$

and

$$\chi_k = \frac{2k\pi i}{\log 2} \quad \text{and} \quad c_k = \frac{1}{\log 2}\Gamma(\chi_k)\zeta(\chi_k)$$

(Γ is Euler's gamma function, ζ is Riemann's zeta funtion).

Let us mention the following expansion ([10])

$$[z^n]\log(1-z)(1-z)^\alpha \sim \frac{-n^{-\alpha-1}\log n}{\Gamma(-\alpha)} + n^{-\alpha-1}\frac{\Gamma(-\alpha)}{(\Gamma(-\alpha))^2},$$

which gives us as a special instance

$$[z^n]\,\varepsilon\log\varepsilon \sim \frac12 4^n n^{-3/2}\left(\frac{\log n}{2\sqrt{\pi}} + \frac{\gamma+2\log 2-2}{2\sqrt{\pi}}\right).$$

We will also use the result

$$[z^n]\,\varepsilon^{1-\chi_k} \sim 4^n n^{(\chi_k-3)/2}\Big/\Gamma\left(\frac{\chi_k-1}{2}\right).$$

We observe that we can simplify the expression

$$c_k\Big/\Gamma\left(\frac{\chi_k-1}{2}\right) = \frac{\zeta(\chi_k)\Gamma\left(\frac{\chi_k}{2}\right)(\chi_k-1)}{4\sqrt{\pi}\log 2};$$

for this we use the duplication formula of *Legendre-Gauss* (see e.g. [2])

$$\Gamma\left(\frac{z}{2}\right)\Gamma\left(\frac{z+1}{2}\right) = \sqrt{\pi}2^{1-z}\Gamma(z).$$

Finally, if we collect everything, we find that

$$\begin{aligned}[z^n]\,E(z) \sim{}& \frac{4^n n^{-3/2}}{2\sqrt{\pi}\log 2}(\log n + \gamma + 2\log 2 - 2)\\ &- A\frac{4^n}{2\sqrt{\pi}}n^{-3/2} + \sum_{k\neq 0}\frac{4^n n^{(\chi_k-3)/2}\zeta(\chi_k)\Gamma(\chi_k/2)(\chi_k-1)}{\sqrt{\pi}\log 2}.\end{aligned}$$

If we take into consideration that

$$b_n \sim \frac{4^n}{\sqrt{\pi}}n^{-3/2} \qquad \text{and} \qquad n^{\chi_k/2} = \varepsilon^{\log_4 n\cdot 2k\pi i},$$

we obtain

Theorem 1. *The average number R_n of registers to evaluate a binary tree of size n verifies*

$$R_n = \log_4 n + D(\log_4 n) + o(1),$$

where $D(x)$ is a continuous periodic function of period 1; this function can be expanded in a Fourier series as

$$D(x) = \sum_{k \in \mathbb{Z}} d_k e^{2k\pi i x},$$

with

$$d_0 = \frac{1}{2} - \frac{\gamma}{2\log 2} - \frac{1}{\log 2} + \log_2 \pi$$

and

$$d_k = \frac{1}{\log 2}\zeta(\chi_k)\Gamma(\frac{\chi_k}{2})(\chi_k - 1), \quad k \neq 0, \quad \chi_k = \frac{2k\pi i}{\log 2}.$$

The remainder of this section ist devoted to the study of the expansion of the function $E(z)$.

To study $E(z)$ for $z \to \frac{1}{4}$, i.e. for $u \to 1$, we set $u = e^{-t}$ and consider $E(z)$ for $t \to 0$. We easily obtain that

$$\frac{1-u^2}{u} \sim 2t + \dots\ .$$

Let us turn to the second factor of $E(z)$, which we denote by $V(t)$; the Mellin transform of $V(t)$ is denoted by $V^*(s)$:

$$V(t) = \sum_{k \geq 1} v_2(k) e^{-kt}; \qquad V^*(s) = \int_0^\infty x^{s-1} V(x) dx$$

$$\begin{aligned} V^*(s) &= \int_0^\infty x^{s-1} \sum_{k \geq 1} v_2(k) e^{-kx} dx \\ &= \sum_{k \geq 1} v_2(k) k^{-s} \int_0^\infty x^{s-1} e^{-x} dx \end{aligned}$$

We have

$$\sum_{k \geq 1} v_2(k) k^{-s} = 2^{-s} \sum_{k \geq 1} v_2(2k) k^{-s} = 2^{-s} \sum_{k \geq 1} \left[1 + v_2(k)\right] k^{-s},$$

and thus

$$V^*(s) = \frac{\Gamma(s)\zeta(s)}{2^s - 1}.$$

The Mellin inversion formula gives us

$$V(t) = \frac{1}{2\pi i} \int_{2-i\infty}^{2+i\infty} V^*(s) t^{-s} ds.$$

We can shift the line of integration to the left, provided we take the residues of the integrand $V^*(s)t^{-s}$ into account. There are poles at $s = 1$ (originating from the Riemann ζ- function) , at $s = 0, -1, -2, \ldots$ (coming from the function Γ) and at $s = \chi_k$, $k \in \mathbb{Z}$ (from $(2^s - 1)^{-1}$). We use the classical expansions to get

$$V(t) \sim \frac{1}{t} + \frac{1}{2}\log_2 t - \frac{1}{2}\log_2 2\pi + \frac{1}{4} + \frac{\gamma}{2\log 2} + \frac{1}{\log 2}\sum_{k\neq 0}\Gamma(\chi_k)\zeta(\chi_k)t^{-\chi_k} + \ldots$$

which gives us the announcedc formula for $E(z)$, after having multiplied the last 2 expansions and noting that

$$t \sim 2\varepsilon + \ldots .$$

3. A similar problem: odd-even merge

In the analysis of a sorting algorithm (odd-even merge) one has to consider the quantity

$$B_n = \sum_{k\geq 1} \frac{\binom{2n}{n-k}}{\binom{2n}{n}}\Big(2F(k) + k\Big),$$

representing the number of exchanges in the odd-even merge of $2n$ elements. In this equation,

$$F(k) = \sum_{0\leq j<k} f(j)$$

and $f(j)$ is the number of digits "1" in the Gray code representation of j (see [6], [21]). Below we see the first few numbers in this representation:

0	0	1100	8
1	1	1101	9
11	2	1111	10
10	3	1110	11
110	4	1010	12
111	5	1011	13
101	6	1001	14
100	7	1000	15
		11000	16

The numbers

$$\vartheta(j) := f(j) - f(j-1)$$

equal ± 1: if k is of the form $k = 2^m(4i+1)$, $\vartheta(k) = 1$; if k is of the form $k = 2^m(4i+3)$, $\vartheta(k) = -1$. We might recognize a great similarity between $\vartheta(k)$ and

$v_2(k)$ from Section 2. The Dirichlet generating function of the sequence $\vartheta(k)$ is

$$\begin{aligned}\sum_{k\geq 1}\vartheta(k)k^{-s} &= \sum_{m\geq 0}\sum_{i\geq 0}2^{-ms}(4i+1)^{-s} - \sum_{m\geq 0}\sum_{i\geq 0}2^{-ms}(4i+3)^{-s}\\ &= \frac{1}{1-2^{-s}}\cdot\frac{1}{4^s}\left(\zeta\big(s,\frac{1}{4}\big)-\zeta\big(s,\frac{3}{4}\big)\right),\end{aligned}$$

where the Hurwitz $\zeta(s,a)$-function is defined by

$$\zeta(s,a) = \sum_{n\geq 0}(n+a)^{-s} \qquad \text{for } \Re(s)>1.$$

We can write

$$B_n = \frac{n}{2} + \frac{2}{\binom{2n}{n}}C_n,$$

with

$$C_n = \sum_{k\geq 1}F(k)\binom{2n}{n-k}.$$

We want to find an analogous function to $E(z)$ (Section 2). Thus we write

$$\begin{aligned}C_n &= \sum_{k\geq 1}F(k)\,\left[u^{n-k}\right](1+u)^{2n}\\ &= [u^n]\sum_{k\geq 1}F(k)u^k(1+u)^{2n}\\ &= [u^n]\sum_{k\geq 1}\sum_{0\leq j<k}f(j)\;u^k(1+u)^{2n}\\ &= [u^n]\sum_{j\geq 0}f(j)u^j\frac{u}{1-u}(1+u)^{2n}\\ &= [u^n]\sum_{j\geq 0}\sum_{0\leq i\leq j}\vartheta(i)u^j\frac{u}{1-u}(1+u)^{2n}\\ &= [u^n]\sum_{i\geq 0}\vartheta(i)u^i\frac{u}{(1-u)^2}(1+u)^{2n}.\end{aligned}$$

In general, if we have a series $\Psi(z)$, and if we set $z = u/(1+u)^2$, we have by Cauchy's theorem

$$\begin{aligned}[z^n]\,\Psi(z) &= \frac{1}{2\pi i}\oint\frac{dz}{z^{n+1}}\Psi(z)\\ &= \frac{1}{2\pi i}\oint\frac{du(1-u)}{(1+u)^3}\frac{(1+u)^{2n+2}}{u^{n+1}}\Psi\Big(\frac{u}{(1+u)^2}\Big)\\ &= [u^n]\,(1-u)(1+u)^{2n-1}\Psi\Big(\frac{u}{(1+u)^2}\Big).\end{aligned}$$

Thus we can write

$$C_n = [u^n]\,\frac{u(1+u)}{(1-u)^3}\sum_{i\ge 1}\vartheta(i)u^i.$$

The function appearing in the right hand side of this equation is the analogue of $E(z)$ of Section 2. We easily find that

$$\frac{u(1+u)}{(1-u)^3} \sim \frac{2}{t^3} + \dots\ .$$

If we set

$$V(t) = \sum_{k\ge 1}\vartheta(k)e^{-kt},$$

then we have

$$V^*(s) = \Gamma(s)\cdot\frac{1}{2^s}\cdot\frac{1}{2^s-1}\left(\zeta(s,\frac{1}{4}) - \zeta(s,\frac{3}{4})\right).$$

As a consequence we find

$$V(t) = \frac{1}{2\pi i}\int_{2-i\infty}^{2+i\infty}\Gamma(s)\cdot\frac{1}{2^s}\cdot\frac{1}{2^s-1}\left(\zeta(s,\frac{1}{4}) - \zeta(s,\frac{3}{4})\right)t^{-s}dt.$$

The difference of the Hurwitz ζ-functions is an entire function; the integrand has poles at $s=\chi_k$ and $s=0,-1,-2,\dots$. We have (see [22])

$$\zeta(s,\frac{1}{4}) - \zeta(s,\frac{3}{4}) \sim \frac{1}{2} + s\left(2\log\Gamma(\frac{1}{4}) - \frac{1}{2}\log 2 - \log\pi\right) + \dots$$

for $s\to 0$. Also,

$$\begin{aligned}\Gamma(s) &\sim \frac{1}{s} - \gamma + \dots\\ \frac{1}{2^s-1} &\sim \frac{1}{\log 2}\cdot\frac{1}{s} - \frac{1}{2} + \dots\\ (2t)^{-s} &\sim 1 - s\cdot\log(2t) + \dots\end{aligned}$$

which gives us the residue at $s=0$

$$-\frac{5}{4} - \frac{\gamma}{2\log 2} + 2\log_2\Gamma(\frac{1}{4}) - \log_2\pi - \frac{1}{2}\log_2 t.$$

The residue at $s=\chi_k$ for $k\neq 0$ is simply

$$\frac{1}{\log 2}\Gamma(\chi_k)\left(\zeta(\chi_k,\frac{1}{4}) - \zeta(\chi_k,\frac{3}{4})\right)t^{-\chi_k}.$$

We can simplify, because

$$\zeta(\chi_k, \frac{3}{4}) = -\zeta(\chi_k, \frac{1}{4}).$$

Thus we obtain

$$\begin{aligned} C_n \sim [z^n] \frac{2}{t^3} \Big(& -\frac{1}{2}\log_2 t - \frac{5}{4} - \frac{\gamma}{2\log 2} + 2\log_2 \Gamma(\frac{1}{4}) - \log_2 \pi \\ & + \frac{2}{\log 2} \sum_{k\neq 0} \Gamma(\chi_k)\zeta(\chi_k, \frac{1}{4}) t^{-\chi_k} \Big), \end{aligned}$$

where, with $t \sim 2\varepsilon$

$$\begin{aligned} C_n \sim [z^n] \Big(& -\frac{1}{8}\varepsilon^{-3}\log_2 \varepsilon - \varepsilon^3 \Big(\frac{7}{16} + \frac{\gamma}{8\log 2} - \frac{1}{2}\log_2 \Gamma(\frac{1}{4}) + \frac{1}{4}\log_2 \pi \Big) \\ & + \frac{1}{2\log 2} \sum_{k\neq 0} \Gamma(\chi_k)\zeta(\chi_k, \frac{1}{4}) \varepsilon^{-3-\chi_k} \Big). \end{aligned}$$

We find, as in Section 2,

$$\begin{aligned} [z^n]\, \varepsilon^3 \log \varepsilon &\sim \frac{4^n\sqrt{n}}{\sqrt{\pi}} \Big(-\log n + (2 - \gamma - 2\log 2) \Big), \\ [z^n]\, \varepsilon^{-3} &\sim \frac{2 \cdot 4^n \sqrt{n}}{\sqrt{\pi}}, \\ [z^n]\, \varepsilon^{-3-\chi_k} &\sim 4^n \frac{n^{(1+\chi_k)/2}}{\Gamma\left(\frac{3+\chi_k}{2}\right)}. \end{aligned}$$

By regrouping the preceeding results we obtain

$$\begin{aligned} C_n \sim & \frac{4^n\sqrt{n}}{\sqrt{\pi}} \log n \cdot \frac{1}{8\log 2} \\ & + \frac{4^n\sqrt{n}}{\sqrt{\pi}} \Big(-\frac{1}{4\log 2} - \frac{\gamma}{8\log 2} - \frac{5}{8} + \log_2 \Gamma(\frac{1}{4}) - \frac{1}{2}\log_2 \pi \Big) \\ & + \frac{1}{2} 4^n \sqrt{n} \frac{1}{\log 2} \sum_{k\neq 0} \Gamma(\chi_k)\zeta(\chi_k, \frac{1}{4}) \frac{n^{\chi_k/2}}{\Gamma\left(\frac{3+\chi_k}{2}\right)}. \end{aligned}$$

We have by the duplication formula

$$\Gamma(\chi_k) \Big/ \Gamma\left(\frac{3+\chi_k}{2}\right) = \frac{1}{1+\chi_k} \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{\chi_k}{2}\right).$$

With

$$\binom{2n}{n} \sim \frac{4^n n^{-1/2}}{\sqrt{\pi}}$$

we find

$$\begin{aligned}\frac{D_n}{\binom{2n}{n}}\cdot\frac{1}{n} \sim \frac{1}{8}\log_2 n \\ &- \frac{1}{4\log 2} - \frac{\gamma}{8\log 2} - \frac{5}{8} + \log_2\Gamma\big(\frac{1}{4}\big) - \frac{1}{2}\log_2\pi \\ &+ \frac{1}{2\log 2}\sum_{k\neq 0}\zeta\big(\chi_k,\frac{1}{4}\big)\frac{\Gamma(\chi_k/2)}{1+\chi_k}n^{\chi_k/2}.\end{aligned}$$

Finally we have obtained the

Theorem 2. *The average number of exchanges in the odd-even merge of $2n$ elements verifies*

$$B_n \sim \frac{1}{4}n\log_2 n + nB\big(\log_4 n\big),$$

where $B(x)$ is a continuous periodic function of period 1; this function can be expanded as a Fourier series

$$B(x) = \sum_{k\in\mathbb{Z}} b_k e^{2k\pi i x},$$

with

$$b_0 = -\frac{1}{2\log 2} - \frac{\gamma}{4\log 2} - \frac{3}{4} + 2\log_2\Gamma\big(\frac{1}{4}\big) - \log_2\pi$$

and for $k \neq 0$

$$b_k = \frac{1}{\log 2}\zeta\big(\chi_k,\frac{1}{4}\big)\frac{\Gamma(\chi_k/2)}{1+\chi_k}.$$

4. The height of leaves in planted plane trees

We can understand the problem of this section most easily by considering the following example of a planted plane tree:

This tree has 4 leaves of height (from left to right) $3, 3, 4, 2$. We remark that the height of the tree is given by the height of the highest leaf We imagine now that the leaves are ordered with respect to their heights. In our example, the sequence of heights of leaves is $4, 3, 3, 2$. Let us remark that there might be more leaves with the same height. To simplify the computations, we complete this sequence by an infinity of zeroes. Now it makes sense to speak about the d-th highest leaf for each d, even if the tree has less than d leaves.

Now we consider $h_n^{(d)}$, the average height of the d-th highest leaf where we remark again that $h_n^{(1)}$ is simply the average height of a tree of size n.

Some technical lemmas have appeared already in several papers, see [1], [12], [13], [14], [20].

Lemma 1. *Let $F(z,y)$ be the generating function where $[z^n y^m]\,F$ is the number of trees of size n with m leaves (the tree $\circ$ has zero leaves); we have for each $p \geq 0$*

$$F^p(z,y) = z^p + p \sum_{0 \leq \lambda < n} z^n y^{n-\lambda} \frac{1}{n} \binom{n}{\lambda} \binom{n-p-1}{\lambda-p}.$$

Idea of proof. (See [12]) We introduce $G(z,y)$, the analogous quantity, where $\circ$ has a leaf; translating the symbolic equation for the family of planted plane trees as usual we obtain

$$G(z,y) = z\Big(y + \sum_{r \geq 1} G^r(z,y)\Big) = \big(1 - z(1-y)\big) P\left(\frac{zy}{\big(1-z(1-y)\big)^2}\right).$$

We use the classical expansion

$$P^m(z) = \sum_{k \geq 0} \left[\binom{2k+m-1}{k} - \binom{2k+m-1}{k-1}\right] z^{k+m}.$$

After this we use the binomial theorem 3 times and use the identity

$$\sum_{k \geq 0} \binom{r}{k} \binom{s+k}{p} (-1)^k = (-1)^r \binom{s}{p-r}$$

to obtain

$$G^p(z,y) = z^p y^p + p \sum_{p \leq \lambda \leq n} z^n y^{\lambda} \frac{1}{n} \binom{n}{\lambda} \binom{n-p-1}{\lambda-p}.$$

Finally we can observe that

$$F(z,y) = G(z,y) - zy + z = G(zy, y^{-1}),$$

and obtain thus the announced formula.

Lemma 2. *Let $A_k(z)$ be the generating function of trees of height $\le k$ and set, as before, $z = u/(1+u)^2$. We have*

$$A_k(z) = \frac{u}{1+u} \cdot \frac{1-u^k}{1-u^{k+1}}.$$

Proof. Translating the symbolic equation for the family $\mathcal{P}$, we obtain

$$A_k = \frac{z}{1-A_{k-1}}, \qquad k \ge 1, \quad A_0 = 0;$$

from this we can prove the formula easily by induction . $\square$

Lemma 3. *Let $Q_k(z,y)$ be the generating function, where $[z^n y^m]Q_k$ is the number of trees of size n with m nodes at level k, which is the maximal level. (This means that, if $m \ge 1$, the height of the tree is exactly k.) We have for $k \ge 1$ and $d \ge 1$*

$$[y^d]Q_k(z,y) = A_{k-1}^{d+1}(z) - A_{k-1}^{d}(z) + zA_{k-1}^{d-1}(z).$$

Proof. Translating as usual, we obtain

$$Q_1 = zy \qquad \text{and} \qquad Q_k = \frac{z}{1-Q_{k-1}} \quad \text{for } k \ge 2.$$

Since we have

$$[y^0]Q_k(z,y) = A_{k-1}(z),$$

it is sufficient to show that

$$\begin{aligned} Q_k(z,y) &= A_{k-1}(z) + \sum_{d\ge 1} y^d \Big(A_{k-1}^{d+1}(z) - A_{k-1}^{d}(z) + zA_{k-1}^{d-1}(z) \Big) \\ &= \frac{A_{k-1}}{1-yA_{k-1}} - \frac{yA_{k-1}}{1-yA_{k-1}} + \frac{zy}{1-yA_{k-1}} \\ &= \frac{(1-y)A_{k-1}+zy}{1-yA_{k-1}}; \end{aligned}$$

this is a simple exercise, using the recurrence for the A_k's.

Let $B_{d,k}(z)$ be the generating function where $[z^n]B_{d,k}$ is the number of trees of size n with d leaves having a height $> k$ and all the others having a height $\le k$. We can observe that $B_{0,k} + B_{1,k} + \cdots + B_{d-1,k}$ is the generating function of trees where d-th highest leaf (in the sense described above) has a height $\le k$. Thus the quantity $h_n(d)$ is given by

$$h_n^{(d)} = \frac{[z^n] \sum_{k\ge 0} \Big(P(z) - B_{0,k}(z) - B_{1,k}(z) - \cdots - B_{d-1,k}(z) \Big)}{[z^n]\, P(z)}.$$

Lemma 4. *The average height of a planted plane tree of size n verifies*

$$\frac{[z^n]\sum_{k\ge 0}\Big(P(z)-B_{0,k}(z)\Big)}{[z^n]\,P(z)} \sim \sqrt{\pi n}-\frac{1}{2}+\dots\ .$$

Idea of proof. We find that

$$\begin{aligned}\sum_{k\ge 0}\Big(P(z)-B_{0,k}(z)\Big) &= \sum_{k\ge 0}\Big(P(z)-A_k(z)\Big)\\ &= \frac{1-u}{1+u}\sum_{k\ge 1}\frac{u^k}{1-u^k} = \frac{1-u}{1+u}\sum_{k\ge 1} d(k)u^k\end{aligned}$$

where $d(k)$ is the number of divisors of k. We observe that

$$\sum_{k\ge 1} d(k)k^{-s} = \zeta^2(s),$$

and use the Mellin transform as in Section 2.

As is the Sections 2 and 3 we are going to show that

$$\sum_{k\ge 0} B_{s+1,k}(z) \sim A_s - B_s t+\dots,$$

where A_s and B_s are constants and $u=e^{-t}$. Since

$$P(z)=\frac{u}{1+u}\sim\frac{1}{2}-\frac{1}{4}t+\dots,$$

we can conclude that

$$[z^n]\sum_{k\ge 0} B_{s+1,k}(z)\Big/ t_n \sim 4B_s.$$

Lemma 5. *We have for $s\ge 0$*

$$\sum_{n\ge p} z^n\binom{n}{s}\binom{n-p}{s} = \sum_{j=0}^{s}\frac{z^{s+p+j}}{(1-z)^{s+1+j}}\,\frac{(p+s)!(s+j)!}{j!(s-j)!(p+j)!s!}.$$

Proof. With $D=d/dz$ and $(n)_d = n(n-1)\dots(n-d+1)$ we have

$$\begin{aligned}\sum_{n\ge p} z^n\binom{n}{s}\binom{n-p}{s} &= \frac{1}{(s!)^2}\sum_{n\ge 0}(n)_s z^s D^s z^{n+p}\\ &= \frac{1}{(s!)^2} z^s D^s z^{p+s}\sum_{n\ge 0} z^n\\ &= \frac{1}{s!} z^s D^s \frac{z^{p+s}}{(1-z)^{s+1}}\\ &= \frac{1}{s!} z^s \sum_{j=0}^{s}\binom{s}{j}\big[D^{s-j} z^{p+s}\big]\big[D^j(1-z)^{-s-1}\big]. \quad \square\end{aligned}$$

Lemma 6. *We have for $s \ge 0$ and $k \ge 1$*

$$B_{s+1,k}(z) = \sum_{j=0}^{s} \frac{(s+j)!}{j!(s-j)!(s+1)!} \left(\frac{u}{1+u+u^2}\right)^{s+1+j} \\ \times \frac{(1-u)^2}{1+u} \frac{u^{k-1}}{(1-u^{k-1})(1-u^k)} \sum_{p\ge1} \frac{p(p+s)!}{(p+j)!} A_{k-1}^p.$$

Proof. We replace each leaf at maximal level k (marked by y in $Q_k(z,y)$) by an arbitrary tree whose leaves are marked by y. Now

$$\begin{aligned} B_{s+1,k} &= [y^{s+1}]\, Q_k\Big(z, \frac{1}{z}F(z,y)\Big) \\ &= [y^{s+1}] \sum_{p\ge1} \Big[\frac{1}{z}F(z,y)\Big]^p \Big[A_{k-1}^{p+1} - A_{k-1}^p + zA_{k-1}^{p-1}\Big] \\ &= \sum_{p\ge1} z^{-p} p \sum_{n\ge1} z^n \frac{1}{n}\binom{n}{s+1}\binom{n-p-1}{s} A_{k-1}^p \big[A_{k-1} - 1 + zA_{k-1}^{-1}\big]. \end{aligned}$$

The announced formula is now obtained by remarking that

$$A_{k-1} - 1 + zA_{k-1}^{-1} = \frac{u^{k-1}}{(1-u^{k-1})(1-u^k)} \frac{(1-u)^2}{1+u},$$

$$\frac{z}{1-z} = \frac{u}{1+u+u^2}$$

and using Lemma 5.

Lemma 7. *The generating function $B_{s+1,k}(z)$ is given by*

$$B_{s+1,k}(z) = \sum_{j=0}^{s}\sum_{i=0}^{s-j} \frac{(s+j)!(i+1)}{j!(s-j-i)!(j+i+1)!} \times \\ \times \left(\frac{u}{1+u+u^2}\right)^{s+1+j} (1-u)^2 u^{i+1} u^{k-1} \frac{(1-u^{k-1})^i}{(1-u^{k+1})^{i+2}}.$$

Proof. We use the formula (being a consequence of the *Vandermonde* formula)

$$\frac{p(p+s)!}{(s-j)!(s+1)!(p+j)!} = \sum_{i=0}^{s-j} (i+1)\binom{p}{i+1} \frac{1}{(s-j-i)!(j+i+1)!}$$

and also the equations

$$\frac{A_k}{z} = \frac{1}{1-A_{k-1}}$$

and

$$\sum_{p\ge1} \binom{p}{i+1} A_{k-1}^p = \frac{A_{k-1}^{i+1}}{(1-A_{k-1})^{i+2}}.$$

Now we are ready to attack the local expansion of $\sum_{k\ge0} B_{s+1,k}$. We easily see that

$$\frac{u}{1+u+u^2} \sim \frac{1}{3}; \qquad (1-u)^2 \sim t^2(1-t); \qquad u^{i+1} \sim 1-(i+1)t.$$

Lemma 8. *We have the local expansion*

$$\sum_{k\ge 0} u^{k-1}\frac{(1-u^{k-1})^i}{(1-u^{k+1})^{i+2}}$$
$$\sim t^{-2}(1+2t)\Bigg(\sum_{l=0}^{i}\binom{i}{l}(-2)^l(1+lt)\Big(\zeta(l+2)-\frac{l}{2}\zeta(l+1)t\Big)-\frac{1}{2}t\Bigg).$$

Proof. We can write

$$\begin{aligned}\sum_{k\ge 0} u^{k-1}\frac{(1-u^{k-1})^i}{(1-u^{k+1})^{i+2}} &= u^{-2}\sum_{k\ge 1}\frac{u^k}{(1-u^k)^2}\left(1+\frac{(1-u^{-2})u^k}{1-u^k}\right)^i\\ &= u^{-2}\sum_{l=0}^{i}\binom{i}{l}\left(1-u^{-2}\right)^l\sum_{k\ge 1}\frac{u^{k(l+1)}}{(1-u^k)^{l+2}}.\end{aligned}$$

It is easy to see that

$$u^{-2}\sim 1+2t \qquad \text{and} \qquad (1-u^{-2})^l\sim(-2)^l t^l(1+lt).$$

We use again the Mellin transform:

$$\begin{aligned}\mathcal{M}\sum_{k\ge 1}\frac{e^{-tk(l+1)}}{(1-e^{tk})^{l+2}} &= \sum_{k\ge 1}k^{-s}\mathcal{M}\frac{e^{-t(l+1)}}{(1-e^{t})^{l+2}}\\ &= \zeta(s)\Gamma(s)\sum_{\lambda\ge 0}\binom{\lambda}{l+1}\lambda^{-s}\\ &= \zeta(s)\Gamma(s)\sum_{\lambda\ge 0}\frac{1}{(l+1)!}\sum_{i=0}^{l+1}s_{l+1,i}\lambda^i\lambda^{-s}\end{aligned}$$

(where the $s_{l+1,j}$ are *Stirling* numbers of the first kind)

$$=\frac{1}{(l+1)!}\zeta(s)\Gamma(s)\sum_{i=0}^{l+1}s_{l+1,i}\zeta(s-i).$$

As in Secion 2 we use the Mellin inversion formula; change the line of integration and take the residues into account. In our problem is sufficient to consider the terms where $i=s+1$ and $i=s$. We note that

$$s_{l+1,l+1}=1 \qquad \text{and} \qquad s_{l+1,l}=-\binom{l+1}{2}.$$

Lemma 9. *The sum of the $B_{s+1,k}(z)$ has to local expansion*

$$\sum_{k\geq 0} B_{s+1,k}(z) \sim const - \frac{1}{2}t\sum_{j=0}^{s}\sum_{i=0}^{s-j}\frac{(s+j)!(i+1)}{j!(s-j-i)!(j+i+1)!}3^{-s-1-j}.$$

Idea of proof.

$$\sum_{k\geq 0} B_{s+1,k}(z) \sim \sum_{j=0}^{s}\sum_{i=0}^{s-j}\frac{(s+j)!(i+1)}{j!(s-j-i)!(j+i+1)!}3^{-s-1-j}t^2(1-t)\big(1-(i+1)t\big)\times$$
$$\times\, t^{-2}(1+2t)\Bigg(\sum_{l=0}^{i}\binom{i}{l}(-2)^l(1+lt)\Big(\zeta(l+2)-\frac{l}{2}\zeta(l+1)t\Big)-\frac{1}{2}t\Bigg)$$

We want to obtain the coefficient of t in this expression. This is a simple thing; after some manipulations there are luckily of lot of simplifications . $\Box$

From the preceeding equations we deduce

$$[z^n]\,B_{s+1,k}(z)\Big/t_n \sim 4\cdot\frac{1}{2}\sum_{j=0}^{s}\sum_{i=0}^{s-j}\frac{(s+j)!(i+1)}{j!(s-j-i)!(j+i+1)!}3^{-s-1-j}.$$

For the quantity $h_n^{(d)}$ we have to sum up:

$$\sum_{s=0}^{d-2}2\sum_{j=0}^{s}\sum_{i=0}^{s-j}\frac{(s+j)!(i+1)}{j!(s-j-i)!(j+i+1)!}3^{-s-1-j}.$$

We can rewrite this quantity as

$$\sum_{s=0}^{d-2}\frac{2}{(s+1)3^{s+1}}\,[x^s]\,\frac{1}{(1-x)^2}\left(\frac{1+x}{1-\frac{x}{3}}\right)^{s+1}.$$

Thus we have finally obtained

Theorem 3. *The average height of the d-th highest leaf of a planted plane tree of size n, has the following asymptotic expansion (for d fixed and $n\to\infty$)*

$$h_n^{(d)} = \sqrt{\pi n}-\frac{1}{2}-\sum_{s=0}^{d-2}\frac{2}{(s+1)3^{s+1}}\,[x^s]\,\frac{1}{(1-x)^2}\left(\frac{1+x}{1-\frac{x}{3}}\right)^{s+1} + \ o(1).$$

5. Some remarks concerning the technique of the extraction of coefficients

If one follows the strategy mentioned in the title, the enumeration problems for trees often involve the evaluation of sums of the type

$$\sum_{k\ge 1}\omega(k)\frac{\binom{2n}{n+k+1}-2\binom{2n}{n+k}+\binom{2n}{n+k-1}}{\binom{2n}{n}}.$$

where $\omega(k)$ might be, for example, $v_2(k)$ or $\vartheta(k)$ or $d(k)$ as in Sections 2, 3 and 4.

One needs to approximate certain binomial coefficients in a central region (e.g. $|k|\le\sqrt{n}\log n$). We obtain, for instance by *Stirling*'s formula,

$$\frac{\binom{2n}{n+k}}{\binom{2n}{n}}\sim e^{-k^2/n}\cdot\left(1+\frac{k^2}{2n^2}-\frac{k^4}{6n^3}+\dots\right)$$

(see [9] for a general formula). Quite often r-th differences of binomal coefficients appear, viz.

$$\xi=[u^n]\,u^k(1-u)^r(1+u)^{2n}\Big/\binom{2n}{n}=\sum_{\lambda=0}^{r}(-1)^\lambda\binom{r}{\lambda}\binom{2n}{n+k+r}\Big/\binom{2n}{n}.$$

Let us recall the difference operator Δ:

$$\Delta^r\binom{2n}{n+k}=\sum_{\lambda=0}^{r}\binom{r}{\lambda}(-1)^{r-\lambda}\binom{2n}{n+k+\lambda}$$

Let us consider $\Delta^r e^{-k^2/n}$. For this, we need the *Hermite* polynomials. Here are the necessary properties [3]:

$$e^{2zt-t^2}=\sum_{n\ge 0}H_n(z)\frac{t^n}{n!}$$

$$H_0=1,\quad H_1=2z,\quad H_{n+1}=2zH_n-2nH_{n-1}$$

$$H_n(z)=\sum_{k\ge 0}\frac{n!}{k!(n-2k)!}(-1)^k(2z)^{n-2k}$$

$$H_n(-z)=(-1)^nH_n(z)$$

Thus

$$\begin{aligned}\Delta^r e^{-k^2/n}&=\sum_{\lambda=0}^{r}(-1)^{r-\lambda}\binom{r}{\lambda}e^{-(k+\lambda)^2/n}\\&=e^{-k^2/n}\sum_{\lambda=0}^{r}(-1)^{r-\lambda}\binom{r}{\lambda}e^{-\frac{2k\lambda}{n}-\frac{\lambda^2}{n}}\\&=e^{-k^2/n}\sum_{\lambda=0}^{r}(-1)^{r-\lambda}\binom{r}{\lambda}\sum_{m\ge 0}H_m\left(-\frac{k}{\sqrt{n}}\right)\left(\frac{\lambda}{\sqrt{n}}\right)^m\frac{1}{m!}.\end{aligned}$$

We know (with *Stirling* numbers of the second kind) that

$$\sum_{\lambda=0}^{r}(-1)^{r-\lambda}\binom{r}{\lambda}\lambda^m = r!S_{m,r}.$$

Hence

$$(-1)^r\Delta^r e^{-k^2/n} \sim e^{-k^2/n}\left[H_r\left(\frac{k}{\sqrt{n}}\right)\frac{1}{n^{r/2}} - H_{r+1}\left(\frac{k}{\sqrt{n}}\right)\frac{1}{n^{(r+1)/2}}\frac{r}{2}+\dots\right].$$

Therefore we have

$$\xi \sim e^{-k^2/n}H_r\left(\frac{k}{\sqrt{n}}\right)\frac{1}{n^{r/2}},$$

which was already announced in [5].

To obtain more terms we have to consider $\Delta^r k^t e^{-k^2/n}$.

We can use the general formula (of *Leibniz* type):

$$\Delta^r f(k)g(k) = \sum_{\lambda=0}^{r}\binom{r}{\lambda}\Big(\Delta^{r-\lambda}f(k+\lambda)\Big)\Big(\Delta^{\lambda}g(k)\Big)$$

with $f(k)=k^t$ and $g(k)=e^{-k^2/n}$.

If we have to consider (for a certain polynomial p)

$$[u^n]\,u^k p(u)(1+u)^{2n},$$

we can developp $p(u)$ in powers of $1-u$ and use the just described machinery.

In a lot of other cases we must investigate the quantities

$$[u^n]\,u^k\frac{1}{1-u}(1+u)^{2n};$$

for this we have to consider

$$\sum_{j\ge k}e^{-j^2/n} \sim \sqrt{n}\,\mathrm{erfc}\left(\frac{k}{\sqrt{n}}\right)+\frac{1}{2}e^{-k^2/n}+\dots,$$

which comes for *Euler*'s summation formula and the error function

$$\mathrm{erfc}(z) := \int_z^\infty e^{-t^2}dt.$$

To obtain more terms, we must consider

$$I_m(z) = \int_z^\infty t^m e^{-t^2}dt;$$

we can evaluate this integral by partial integration:

$$I_m(z) = \frac{m-1}{2}I_{m-2}(z)+\frac{1}{2}z^{m-1}e^{-z^2},$$
$$I_1(z) = \frac{1}{2}e^{-z^2}, \qquad I_0(z) = \mathrm{erfc}(z).$$

Finally a more general problem of the same nature appears in the evaluation of sums of the type

$$[u^n]\,u^k\frac{1}{(1-u)^a}(1+u)^{2n}$$

where a is an integer ≥ 1.

References

1. de Bruijn, Knuth, Rice, *The average height of planted plane trees*, Graph Theory and Computing (R.C. Read, eds.), Academic Press, New York, 1972, pp. 15–22.
2. Dieudonne, *Calcul Infinitesimal*, Hermann, Paris, 1968.
3. Erdelyi, Magnus, Oberhettinger, Tricomi, *Higher Transcendental Functions*, Mc-Graw-Hill, 1955.
4. Flajolet, *Analyse d'algorithmes d'arbres and de fichiers*, Cahiers du BURO **34–35** (1981), 1–209.
5. Flajolet, Raoult, Vuillemin, *The number of registers required for evauluating arithmetic expressions*, Theor. Comput. Sci. **9** (1979), 99–125.
6. Flajolet, Ramshaw, *A note on Gray Code and odd-even merge*, SIAM J. on Computing **9** (1980), 142–158.
7. Flajolet, Odlyzko, *The average height of binary trees and other simple trees*, J. Comput. Syst. Sci. **25** (1982), 171–213.
8. Flajolet, Prodinger, *Register Allocation for unary-binary Trees*, SIAM J. on Computing **15** (1986), 629–640.
9. Greene, Knuth, *Mathematics for the Analysis of Algorithms*, Birkhäuser, 1981.
10. Jungen, *Sur les series de Taylor n'ayant que des singularites algebro-logarithmiques sur leur cercle de convergence*, Commentarii Math. Helvetici **3** (1931), 266–306.
11. Kemp, *The average number of registers to evaluate a binary tree optimally*, Acta. Inf. **11** (1979), 363–372.
12. Kemp, *On the average oscillation of a stack*, Combinatorica **2** (1982), 157–176.
13. Kemp, *The average height of planted plane trees with m leaves*, Journal of Combinatorial Theory Ser. B **B 34** (1983), 191–208.
14. Kemp, *On the number of deepest nodes in ordered trees*, Discrete Mathematics **81** (1990), 247–258.
15. Knuth, *The Art of Computer Programming, Vol.1*, Addison-Wesley, 1968.
16. Knuth, *The Art of Computer Programming, Vol.3*, Addison-Wesley, 1968.
17. Meir, Moon and Pounder, *On the order of random channel networks*, SIAM J. on Alg. Disc. Meth. **1** (1980), 25–33.
18. Prodinger, *The average height of the second highest leaf of a planted plane trees*, Europ. J. on Combinatorics **5** (1984), 351–357.
19. Prodinger, *Die Bestimmung gewisser Parameter von binären Bäumen mit Hilfe analytischer Methoden*, Lecture Notes in Mathematics (E.Hlawka, Ed.) **1114** (1985), 118–133.
20. Prodinger, *The Average Height of the d-th Highest Leaf of a Planted Plane Tree*, Networks **16** (1985), 67–75.
21. Sedgewick, *Data movement in odd-even merge*, SIAM J. on Computing **7** (1978), 239–272.
22. Whittaker, Watson, *A course in modern analyis*, Cambridge University Press, 1930.