# The Cost Structure of Quadtrees

*Bruno Salvy*
INRIA Rocquencourt

October 4, 1993

[summary by Philippe Dumas and Michèle Soria]

**Abstract**

Many characteristics of quadtrees, like search costs and page occupancy, are precisely analysed. The mean value of such parameters are shown to have generating functions of hypergeometric type. Integral representations and singularity analysis give explicit forms for many structural constants of those trees.

The quadtree structure is a natural generalization of binary search trees to $d$-dimensional data. It constitutes a fundamental hierarchical representation of point data in higher dimensional spaces, which is used in many different fields like data bases or image processing (see e.g. Samet's book [7]).

Interesting parameters in the study of quadtrees are path length (related to the cost of searching or inserting data) and the page occupancy in the case of quadtrees depending on an integer parameter $b$ representing a page capacity. These additive parameters have been largely investigated: for example, the average path length for $d$-dimensional quadtrees of size $n$ is equivalent to $\frac{2}{d}\,n\log n$ [3]; the page occupancy, in the case $d=2$, approaches 33% [5].

The method used by Flajolet *et alii* in [3, 4, 5] relies on studying a linear differential equation of order $d$ and the local behaviour of its solutions. The results presented here rely on a new method of attack, in which the dimension $d$ is only a parameter in some linear differential equation of order 1. With this method, it is possible to get more precise asymptotic expansions for the average value of parameters, and to obtain formal expressions for the coefficients of these expansions for all $d$'s (but goldies). The previous method, in the case $d>2$, only gave an equivalent and the involved constant was not always reachable.

## 1. Classical method

All cost functions which are studied here are additive and the mean value $f_n$ over all $d$-quadtrees of size $n$ satisfies a recurrence

$$f_n = t_n + 2^d \sum_{k=0}^{n} \pi_{n,k}\, f_k, \tag{1}$$

where $t_n$ is a toll function, that is to say the cost of dividing into sub-problems and reconstructing from sub-problems. For example, if $t_n=1$ then $f_n$ is the number of nodes, and if $t_n=n$ then $f_n$ is the average path length. Besides, $\pi_{n,k}$ is the probability that the first subtree has size $k$, which has value

$$\pi_{n,k} = \frac{1}{n}\sum_{k\le N_1\le\cdots\le N_{d-1}\le n-1}\frac{1}{(N_1+1)\cdots(N_{d-1}+1)}.$$

Translating equation (1) into generating functions gives for $f(z) = \sum_{n=0}^{\infty} f_n\, z^n$ the integral equation

$$f(z) = t(z) + 2^d\, J^{d-1} I f(z), \tag{2}$$

where $I$ and $J$ are two linear operators defined by

$$If(z) = \int_0^z \frac{f(t)}{1-t}\, dt, \quad Jf(z) = \int_0^z \frac{f(t)}{t(1-t)}\, dt.$$

The functional equation (2) may be expressed as a linear differential equation of order $d$, namely

$$\left[z(1-z)\frac{d}{dz}\right]^d \{f(z) - t(z)\} = 2^d z f(z). \tag{3}$$

The asymptotic behaviour of the sequence $(f_n)$ depends on the dominant singularities of $f(z)$, which is a solution of (3).

Let us remind that for a linear differential equation

$$a_k(z)\, y^{(k)}(z) + \cdots + a_0(z)\, y(z) = 0,$$

where the $a_i$'s are polynomials, the singular points are at the roots $\alpha$ of $a_k(z)$. Moreover the local behaviour of the solution may have two forms

$$lbreg(z) = \left(1 - \frac{z}{\alpha}\right)^s \sum_{n \geq 0} P_{k,n}\, (\log(\alpha - z)) \left(1 - \frac{z}{\alpha}\right)^{n/q}$$

or

$$lbirreg(z) = lbreg(z) \times \exp\left[P\left(\frac{1}{(1 - z/\alpha)^{1/q}}\right)\right],$$

according to the regular or irregular type of the singular point $\alpha$ [9]. It is possible to compute all these quantities by a method of indeterminate coefficients [8]. In this way, one obtains a basis of singular solutions (the series may be convergent or divergent). But a basis is not sufficient and one must also find the coordinates of the studied solution with respect to the basis.

Example. The generating function of the average path length in dimension $d = 2$ satisfies the linear differential equation [3]

$$z(1-z)^2 P''(z) + (1 - 2\,z)(1-z)P'(z) - 4\,P(z) = \frac{1 + 3\,z}{(1-z)^2}.$$

In this case, it is possible to give an explicit solution in the form of a hypergeometric function, but we neglect this point of view to illustrate the general method.

First we deal with the homogeneous equation. It has two singularities $\alpha = 0, 1$, and we find two solutions

$$f_1(z) = \frac{1}{(1-z)^2}\left[1 - \frac{2}{3}(1-z) + \cdots\right], \quad f_2(z) = (1-z)^2\left[1 + \frac{6}{5}(1-z) \cdots\right].$$

A particular local solution is

$$f_0(z) = \frac{1}{(1-z)^2}\left[\log\frac{1}{1-z} - \frac{2}{3} + \cdots\right].$$

So the general solution

$$f(z) = f_0(z) + c_1\, f_1(z) + c_2\, f_2(z)$$

has the singular behaviour

$$\frac{1}{(1-z)^2}\log\frac{1}{1-z}+\frac{c_1-2/3}{(1-z)^2}+\cdots,$$

hence for the coefficients

$$f_n = n\log n + (c_1+\gamma-5/3)\,n+\cdots. \tag{4}$$

Thus the recurrence gives first a linear differential equation, next the dominant singularity and local behaviour, and eventually the asymptotic behaviour of $f_n$. Alas we cannot compute the coefficient $c_1$, except numerically.

## 2. New method

The new method relies first on Euler transform, which yields a first order recurrence instead of a full history recurrence, and second on analytically continuation of Taylor series.

**2.1. Euler transform.** The Euler transform

$$f^*(z)=\frac{1}{1-z}f\left(\frac{-z}{1-z}\right)$$

is an involution and the associated relation on coefficients is

$$f_n=\sum_{k=0}^{n}(-1)^k\binom{n}{k}f_k^*.$$

To abbreviate, we note $Z=-z/(1-z)$. According to (2), the function $f^*(z)$ satisfies a new functional equation

$$(1-Z)f^*(Z)=(1-Z)t^*(Z)+2^dJ^{d-1}I(1-Z)f^*(Z),$$

which involves two operators, which are much simpler than the preceding ones,

$$I(1-Z)f^*(Z)=-\int_0^Z f^*(u)\,du,\quad\text{and}\quad Jg(Z)=\int_0^Z\frac{g(u)}{u}\,du.$$

Moreover the underlying recurrence is merely of order 1,

$$f_n^*=u_n+\left[1-\left(\frac{2}{n}\right)^d\right]f_{n-1}^* \tag{5}$$

($u_n=t_n^*-t_{n-1}^*$) and it is to be compared with the original recurrence

$$f_n=t_n+2^d\sum_{k=0}^{n}\pi_{n,k}f_k.$$

Example. For the average path length recurrence, (5) is easy to solve because $u_n=\delta_{n,2}-\delta_{n,1}$ is zero for $n\geq 3$, hence

$$f_n^*=\prod_{k=3}^{n}\left(1-\left(\frac{2}{k}\right)^d\right)\text{ for }n\geq 3.$$

As a result $f^*(z)$ is hypergeometric, hence $f(z)$ is hypergeometric too. More precisely we have

$$f(z)=\frac{z}{(1-z)^2}+\frac{z^2}{(1-z)^3}\,{}_{d+1}F_d\left(\begin{matrix}3-\omega_1,\ldots,3-\omega_d,1\\3,\ldots,3\end{matrix}\,\middle|\,z\right),$$

with the classical notation of generalized hypergeometric functions

$$ {}_pF_q\left( \begin{array}{c} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{array} \middle| z \right) = \sum_{n=0}^{\infty} \frac{(a_1)_n \cdots (a_p)_n}{(b_1)_n \ldots (b_q)_n} \frac{z^n}{n!}, $$

and $(a)_n = a(a+1)\cdots(a+n-1)$. The link between $f_n$ and the hypergeometric series comes from the equality

$$ \prod_{k=3}^{n} \frac{k^d - 2^d}{k^d} = \prod_{k} \prod_{\omega^d = 2^d} \frac{k-\omega}{k} = \prod_{\omega} \frac{(3-\omega)(4-\omega)\cdots(n-\omega)}{3 \cdot 4 \cdots n}. $$

**2.2. General theorem.** In the preceding example, the expression of $f_n$ is explicit. This is a general result because $f_n^*$ satisfies

$$ f_n^* = A(n) \sum_{k=2}^{n} \frac{u_k}{A(k)}. $$

For paged trees, we have $u_k = (-1)^{k+1}\binom{n-2}{k-1}$ and for the number of leaves $u_k$ is simply 1. We obtain in this way the following theorem [2].

THEOREM 1. *The expectation $f_n$ of an additive cost function with toll $t_n$ is*

$$ f_n = t_0 + n\left((2^d - 1)t_0 + t_1\right) + \sum_{k=2}^{n} (-1)^k \binom{n}{k} A_k \sum_{j=2}^{k} \frac{t_j^* - t_{j-1}^*}{A_j} $$

*with*

$$ A_n = \prod_{j=3}^{n} \left(1 - \frac{2^j}{j^d}\right), \quad t_j^* = \sum_{k=0}^{j} (-1)^k \binom{j}{k} t_k. $$

But the asymptotic behaviour is not yet known. To get that damned behaviour we search for the behaviour of $f^*(z)$ at $-\infty$, since point $-\infty$ corresponds to point 1 (the dominant singularity of $f(z)$) through Euler transformation. Function $f^*(Z)$ is defined as a power series and we search for an integral representation, which permits an extension to the whole plane. The formula for analytic continuation of Taylor series [6] is

$$ g(-t) = \frac{1}{2i\pi} \int_{c-i\infty}^{c+i\infty} \varphi(-s) t^{-s} \frac{\pi}{\sin \pi s}\, ds $$

if we assume that

$$ g(t) = \sum_{n \geq 0} \varphi(n)\,(-t)^n $$

and $\varphi(s)$ is an analytical function, satisfying some growth conditions.

EXAMPLE. For the path length

$$ f_n^* = \prod_{k=3}^{n} \left(1 - \left(\frac{2}{k}\right)^d\right) = \frac{A(n)}{A(2)} $$

with

$$ A(n) = \frac{1}{n(n-1)} \prod_{\substack{\omega^d = 2^d \\ \omega \neq 2}} \frac{\Gamma(n+1-\omega)}{\Gamma(n+1)} $$

and we use

$$ f^*(-t) = \frac{1}{A(2)} \frac{1}{2i\pi} \int_{-3/2-i\infty}^{-3/2+i\infty} A(-s) t^{-s} \frac{\pi}{\sin \pi s}\, ds, $$

as an extension of our power series $f^*(-t)$. Next we shift the line of integration to the right and collect the residues. There is a pole of order 2 at $s=-1$ and the computation of the residue gives

$$f^*(-t) \underset{t\to\infty}{=} t + \frac{2}{d} t \left[\log t - 1 + \sum_{\omega} \psi(2-\omega) - \psi(2)\right] + O\left(\log t + t^{1-2\cos 2\pi/d}\right),$$

where $\psi = \Gamma'/\Gamma$. Using Euler transform, the preceding expansion, valid in a neighbourhood of $-\infty$, becomes an expansion in the neighbourhood of 1, which yields a more precise expression than the one obtained by the classical method –cf. (4)–

$$f_n = \frac{2}{d} n \log n + n\left(1 - \frac{1}{d} + \frac{2\gamma}{d} + \frac{2}{d} \sum_{\substack{\omega^d=2^d \\ \omega\neq 2}} [\psi(2-\omega) - \psi(2)]\right) + O\left(\log n + n^{1-2\cos 2\pi/d}\right).$$

Example. For the number of leaves the formula is

$$f_n^* = A(n) \sum_{k=2}^{n} \frac{1}{A(k)}.$$

The problem is to make this expression an analytic function of $n$. A first way to do this is to use the series of differences (the term $n-1$ is a correction due to $\lim_k A(k) = 1$),

$$f_n^* = A(n) \left[\sum_{k\geq 2} \left(\frac{1}{A(k)} - \frac{1}{A(k+n-1)}\right) + n - 1\right].$$

A second way is to write

$$f_n^* = \lim_{u\to 1} A(n) \sum_{k\geq 2} \left(\frac{u^k}{A(k)} - \frac{u^{k+n-1}}{A(k+n-1)}\right).$$

In both cases we obtain

$$f_n = n\left[1 - \prod_{\substack{\omega^d=2^d \\ \omega\neq 2}} \Gamma(2-\omega)\left(1 + \sum_{k} \frac{A'(k)}{A(k)}\right)\right] + \ldots$$

For $d=2$, the factor of $n$ has value $4\pi^2 - 39 \simeq 0.47841762$.

Example. Eventually the page occupancy for quadtrees with page capacity $b$ gives rise to the sequence

$$f_n^* = A(n) \sum_{k=b+1}^{n} \frac{\binom{k-2}{b-1}}{A(k)}.$$

The continuation of $f_n^*$ as an analytical function of $n$ is more subtle. The first way needs the $b$ first terms of the asymptotic expansion of $f_n^*$ to be precomputed, and this is not satisfactory. The second way uses

$$f_n^* = \lim_{v\to 1} \left\{ A(n) \sum_{j\geq b+1} \left[\binom{j-2}{b-1} \frac{v^j}{A(j)} - \frac{\Gamma(j+n-2)}{(b-1)!\Gamma(j+n-b-1)} \frac{v^{j+n-1}}{A(j+n-1)}\right] \right.$$
$$\left. - A(n) \sum_{j=2}^{b} \frac{\Gamma(j+n-2)}{(b-1)!\Gamma(j+n-b-1)} \frac{v^{j+n-1}}{A(j+n-1)} \right\}.$$

It is noteworthy that the bounds of the sums are independent of $n$. So that it is possible to formally compute the constants in the asymptotic expansion given by the theorem. By the former method, these constants were attainable only by numerical computation.

## 3. Conclusion

This new method, which treats the dimension $d$ as a parameter, permits to study precisely the additive characteristics of quadtrees: full asymptotic expansions are available, coefficients of these expansions are formally computable. But the computation is rather difficult and may involve summation of multiple series. Moreover, only additive parameters can be dealt with. (Such an important parameter as the height must be tackled with other methods [1].) Still this method has the advantage of generality and may be applied to a wide class of problems.

## Bibliography

[1] Devroye (Luc) and Laforest (Louise). – An analysis of random $d$–dimensional quad trees. *SIAM Journal on Computing*, vol. 19, 1990, pp. 821–832.

[2] Flajolet (Ph.), Labelle (G.), Laforest (L.), and Salvy (B.). – *The Cost Structure of Quadtrees.* – Technical Report n° 2249, Institut National de Recherche en Informatique et en Automatique, April 1994.

[3] Flajolet (Philippe), Gonnet (Gaston), Puech (Claude), and Robson (J. M.). – The analysis of multidimensional searching in quad–trees. In *Proceedings of the Second Annual ACM–SIAM Symposium on Discrete Algorithms.* pp. 100–109. – Philadelphia, 1991.

[4] Flajolet (Philippe) and Lafforgue (Thomas). – Search costs in quadtrees and singularity perturbation asymptotics. *Discrete and Computational Geometry*, vol. 12, n° 4, 1994.

[5] Hoshi (Mamoru) and Flajolet (Philippe). – Page usage in a quadtree index. *BIT*, vol. 32, 1992, pp. 384–402.

[6] Lindelöf (Ernst). – Le calcul des résidus et ses applications à la théorie des fonctions. In *Collection de monographies sur la théorie des fonctions, publiée sous la direction de M. Émile Borel.* – Gauthier-Villars, Paris, 1905. Reprinted by J. Gabay, 1989.

[7] Samet (Hanan). – *The Design and Analysis of Spatial Data Structures.* – Addison–Wesley, 1990.

[8] Tournier (Évelyne). – *Solutions formelles d'équations différentielles.* – Doctorat d'État, Université scientifique, technologique et médicale de Grenoble, 1987.

[9] Wasow (W.). – *Asymptotic Expansions for Ordinary Differential Equations.* – Dover, 1987. A reprint of the John Wiley edition, 1965.