22

Limit Distributions in Quadtrees

Thomas Lafforgue
LRI, Orsay

[résumé par Philippe Flajolet]

Quadtrees constitute a classical data structure for storing and accessing multidimensional data. It is proved here that, in all dimensions, the cost of a random search in a randomly grown quadtree has logarithmic mean and variance and that it is asymptotically distributed as a normal variate. The limit distribution property extends to quadtrees a result only known so far to hold for binary search trees.

The analysis is based on a technique of singularity perturbation analysis applied to linear differential equations satisfied by intervening bivariate generating functions.

The work described is based on a joint paper of Lafforgue and Flajolet [4].

1 Introduction

Quadtrees are a well known data structure discovered by Finkel and Bentley which is discussed in classical treatises on algorithms [5, 11] and examined in great detail in Samet's reference books [9, 10]. Their analysis has made visible progress over recent years [2, 3, 6, 7, 8].

The probabilistic model considered takes all data uniformly from the d–dimensional hypercube $\mathcal{Q} = [0,1]^d$. The search cost D_n is defined as the cost—measured in internal nodes traversed—of searching a random point in a randomly grown quadtree of size $n-1$; it is also called the insertion depth of the nth node and is a random variable defined on the space $\mathcal{Q}^{n-1} \times \mathcal{Q} \cong \mathcal{Q}^n$.

The main result reported is that D_n *converges in distribution to a Gaussian variate* when the size $n-1$ of the tree structure becomes large. Figure 1 illustrates the clear occurrence of this phenomenon already for low values of n.

Theorem 1 *(i). The mean μ_n and standard deviation σ_n of the cost D_n of a random search in random quadtree of size $n-1$ in dimension $d \geq 1$ satisfy asymptotically*

$$\mu_n \sim \frac{2}{d}\log n \qquad \text{and} \qquad \sigma_n \sim \sqrt{\frac{2}{d^2}\log n}. \tag{1}$$

(ii). The distribution of D_n converges in distribution to a normal variate: for all real α, β,

$$\Pr\{\alpha \leq \frac{D_n - \mu_n}{\sigma_n} \leq \beta\} \;\to\; \frac{1}{\sqrt{2\pi}}\int_\alpha^\beta e^{-x^2/2}\,dx \qquad (n \to \infty). \tag{2}$$

The mean was determined by Devroye and Laforest [2] and independently by Flajolet, Gonnet, Puech, and Robson [3].

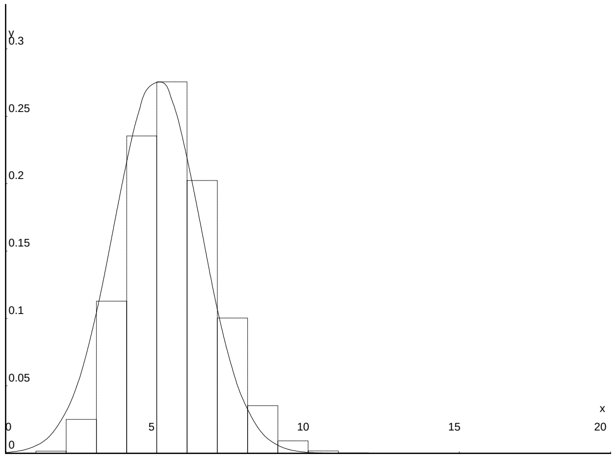

Figure 1: The histogram of the probability distribution of D_n (for dimension $d = 2$ and $n = 100$) plotted against a Gaussian density function of same mean and variance.

2 Basic equations

Two integral operators play an essential rôle here:

$$\mathbf{I}\, f(z) = \int_0^z f(t)\, \frac{dt}{1-t} \qquad \mathbf{J}\, f(z) = \int_0^z f(t)\, \frac{dt}{t(1-t)}.$$

Lemma 1 *The generating functions of the costs of a random search, successful (C_n) and unsuccessful (D_n), in a quadtree of size n are given by*

$$\left\{ \begin{array}{lcl} \gamma_n(u) & := & \sum_k \Pr\{C_n = k\} u^k = \dfrac{1}{n} \dfrac{u}{2^d u - 1}(\phi_n(u) - 1) \\ \delta_n(u) & := & \sum_k \Pr\{D_n = k\} u^k = \dfrac{1}{2^d u - 1}(\phi_n(u) - \phi_{n-1}(u)), \end{array} \right. \tag{3}$$

where the bivariate generating function

$$\Phi(u,z) = \sum_n \phi_n(u) z^n$$

of the level polynomials $\phi_n(u)$ is characterized by the integral equation

$$\Phi(u,z) = 1 + 2^d u \mathbf{J}^{d-1} \mathbf{I}\, \Phi(u,z). \tag{4}$$

Proof. A quadtree of size n gives rise to a root subtree of size k with probability

$$\pi_{n,k} = \frac{1}{n} \sum_{\mathcal{L}} \frac{1}{(\ell_1 + 1)(\ell_2 + 1) \cdots (\ell_{d-1} + 1)}, \tag{5}$$

where the summation is over $n > \ell_1 \geq \ell_2 \geq \cdots \geq \ell_{d-1} \geq k$. The rest relies on simple combinatorial properties of the "level polynomials" ϕ_n and on translating recurrences into generating function equations. □

3 Lower dimensions

The binary search tree ($d = 1$). The main results are originally due to Hibbard for the mean and Lynch for the whole distribution. See [8]. Related results hold for successful searches. When $d = 1$, the integral equation satisfied by $\Phi(u, z)$ is homogeneous of order 1, and thus solvable by quadratures. We find

$$\Phi(u,z) = \frac{1}{(1-z)^{2u}} \qquad \text{and} \qquad \phi_n(u) = \frac{(2u)\cdot(2u+1)\cdots(2u+n-1)}{n!}. \tag{6}$$

Thus, we have $[u^k]\phi_n(u) = 2^k \left[{n \atop k}\right]/n!$, which involves the Stirling numbers of the first kind ("cycle" Stirling numbers).

Theorem 2 (Hibbard, Lynch) *The cost D_n of a random search in a binary search tree of size $n-1$ has mean and variance given by*

$$\mu_n = 2(H_n - 1) \qquad \sigma_n^2 = 2H_n - 4H_n^{(2)} + 2,$$

and probability distribution

$$\Pr\{D_n = k\} = \frac{2^k}{n\,!}\begin{bmatrix} n-1 \\ k \end{bmatrix}.$$

The standard quadtree ($d = 2$). In the case of dimension $d = 2$, the analytic model of quadtrees can be solved explicitly in terms of hypergeometric functions that are otherwise known to occur in the average case analysis of partial match.

Theorem 3 *The cost D_n of a random search in a standard quadtree of size $n-1$ has a generating function $\gamma_n(u)$ given by*

$$\gamma_n(u^2) \equiv \mathbf{E}\{u^{2D_n}\} = \frac{1}{4u^2-1}\sum_{j=0}^{n}\binom{2u}{j}\binom{2u-1}{j}\binom{2u-2+n-j}{n-j}.$$

Thus the probability distribution of D_n is expressible as a complicated convolution of Stirling numbers.
Proof. The generating function Φ of the level polynomials satisfies

$$\Phi(u,z) = 1 + 2^2 u \int_0^z \frac{dx}{x(1-x)} \int_0^x \Phi(u,t)\,\frac{dt}{1-t}$$

an equation whose solution admits an hypergeometric form:

$$\Phi(u^2; z) = \frac{1}{(1-z)^{2u}}\, F[-2u, 1-2u; 1; z]. \tag{7}$$

where

$$F \equiv F[a,b;c;z] = 1 + \frac{a\cdot b}{c}\frac{z}{1!} + \frac{a(a+1)\cdot b(b+1)}{c(c+1)}\frac{z^2}{2!} + \cdots. \tag{8}$$

□

4 The singularity perturbation method

The architecture of the proof of the main theorem asserting asymptotic normality of the distribution is transparent although implementation of it requires quite some care. In essence, we need to solve a double inversion problem in order to recover the coefficients $[z^n u^k]\Phi(u,z)$. A first stage consists in extracting $\phi_n(u) = [z^n]\Phi(u,z)$. This involves examining the influence of a parameter (u) on the singularity (at $z = 1$) of a differential equation. For that reason we call our method a *singularity perturbation method.* It is intermediate in difficulty between regular perturbations and singular perturbations, the latter implying reduction of order[6]. A second stage (from $\phi_n(u)$ to its coefficients) relies on continuity theorems of analytic probability. We offer here a brief outline.

The starting point is the integral equation furnished by Lemma 1,

$$\Phi(u,z) = 1 + 2^d u \mathbf{J}^{d-1}\mathbf{I}\,\Phi(u,z). \tag{9}$$

That equation is itself equivalent to a linear differential equation with coefficients that are polynomial in the main variable z and the parameter u. The order of the equation is equal to the dimension of the data space, d. The standard theory is more conveniently developed from *differential systems* rather than equations, and the associated system is also of dimension d.

The most common case for linear differential equations and systems is the one called *regular singularity* or *singularity of the first kind* [1, 12]. In such a case, a basis of solutions can be found that are of the approximate form $c/(1-z)^\alpha$. The possible exponents α are determined by substituting into the equation. They thus appear as roots of a polynomial called the *indicial polynomial.*

In a parameterized case like (9), we thus expect solutions to involve linear combinations of terms of the form

$$\frac{c(u)}{(1-z)^{\alpha(u)}}, \tag{10}$$

as $z \to 1$. In the case of (9), it is found that the exponents are the algebraic functions that are roots of the indicial equation

$$(\alpha(u))^d - 2^d u = 0.$$

Forms belonging to the general type (10) were already encountered when $d = 1$, see Eq. (6), and when $d = 2$, see Eq. (7).

As $z \to 1$, the dominant term in the expansion of $\Phi(u,z)$ is the one corresponding to the root $2u^{1/d}$ which has maximal real part. In particular when the parameter u is close to 1, this is the principal determination of $2\sqrt[d]{u}$. From the shape (10) of singular elements, we thus expect the singular form of Φ to be

$$\Phi(u,z) \approx \frac{c(u)}{(1-z)^{2u^{1/d}}} \qquad (z \to 1), \tag{11}$$

at least for u near 1.

According to the usual principles of singularity analysis, the *dominant singular behaviour* of Φ provides the dominant asymptotic term in its coefficients $\phi_n(u) = [z^n]\Phi(u,z)$. Translating (11) to coefficients, we expect to get, as an approximation of $\phi_n(u)$,

$$\phi_n(u) \approx c(u)\frac{n^{2u^{1/d}-1}}{\Gamma(2u^{1/d})}. \tag{12}$$

[6] We refer to standard texts like [1, 12] for basics on linear differential equations in the complex domain.

Given the approximation (12), values of the polynomial $\phi_n(u)$ are asymptotically known at least for u in a neighbourhood of 1. An inversion problem (the second one after the phase of singularity analysis ensuring the transition from (11) to (12)) is then to be solved. The approximation (12) permits to estimate $\phi_n(e^{i\theta})$, suitably normalized, when θ lies near 0. The Fourier transform of the distribution defined by the coefficients of $\phi_n(u)$ tends to $e^{-\theta^2/2}$, the characteristic function of the Gaussian distribution:

$$\lim_{n\to+\infty} e^{-i\theta a_n/b_n} \frac{f_n(e^{i\theta/b_n})}{f_n(1)} = e^{-\theta^2/2},$$

for some suitably chosen centering constants a_n, b_n (that may be taken equal to the mean μ_n and variance σ_n of the distribution).

Since $\phi_n(u)$ has positive coefficients, the continuity theorem for characteristic functions (or equivalently Fourier transforms of measures) of analytic probability theory applies. This leads to the end result, namely the convergence in distribution to a normal distribution for the coefficients of $\phi_n(u)$ which in turn carries to the distribution of D_n as expressed by the main theorem.

The method is expected to be instrumental for a wide class of recurrences whose generating functions satisfy parameterized linear differential equations.

References

[1] E. A. Coddington and M. Levinson. *Theory of Ordinary Differential Equations.* McGraw-Hill, 1955.

[2] L. Devroye and L. Laforest. An analysis of random d–dimensional quad trees. *SIAM Journal on Computing*, 19:821–832, 1990.

[3] Ph. Flajolet, G. Gonnet, C. Puech, and J. M. Robson. Analytic variations on quadtrees. *Algorithmica*, 1992. 24 pages, to appear.

[4] Ph. Flajolet and Th. Lafforgue. Search costs in quadtrees and singularity perturbation asymptotics. In preparation, 1992.

[5] G. H. Gonnet and R. Baeza-Yates. *Handbook of Algorithms and Data Structures: in Pascal and C.* Addison–Wesley, second edition, 1991.

[6] M. Hoshi and Ph. Flajolet. Page usage in a quadtree index. *BIT*, 32:384–402, 1992.

[7] L. Laforest. Étude des arbres hyperquaternaires. Technical Report 3, LACIM, UQAM, Montreal, November 1990. (Author's PhD Thesis at McGill University).

[8] H. M. Mahmoud. *Evolution of Random Search Trees.* John Wiley, 1992.

[9] H. Samet. *Applications of Spatial Data Structures.* Addison–Wesley, 1990.

[10] H. Samet. *The Design and Analysis of Spatial Data Structures.* Addison–Wesley, 1990.

[11] R. Sedgewick. *Algorithms.* Addison-Wesley, Reading, Mass., second edition, 1988.

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