Travel Inside a "Funny" Complex Differential Equation

Philippe Jacquet
INRIA

December 13, 1993

[summary by Joris van der Hoeven]

Abstract

We consider a functional-differential equation of the form $h_z(z,u) = h(puz,u)h(quz,u)$, where $p,q > 0$ with $p+q = 1$, $h(z,1) = e^z$ and $h(0,u) = 1$. This equation arises when studying the number of phrases in the fundamental parsing algorithm due to Lempel and Ziv. More precisely, it is shown that this problem is equivalent to a problem on digital trees, which reduces to determining the asymptotics of the above equation.

1. The Lempel-Ziv parsing algorithm

The Lempel-Ziv parsing algorithm takes a word as input and partitions it into phrases (blocks) of variable size. In this partition, each new phrase consists of an old phrase, as long as possible, together with a new letter. For instance, the string 11001010001000100 is parsed into (1)(10)(0)(101)(00)(01)(000)(100). Replacing each new phrase by a pointer to the old phrase, together with the new letter, yields a universal data compression algorithm [9]. Other applications are tests of randomness, efficient transmission of data, etc [7, 8, 9].

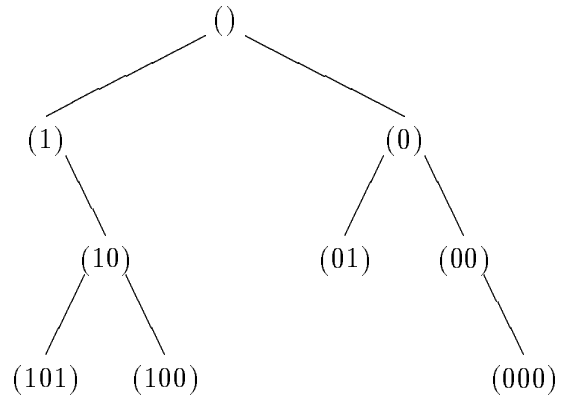

FIGURE 1. The digital tree associated to the string 11001010001000100

Different parameters can be associated to this parsing algorithm, to start with the input size n and the number of phrases $m = M_n$. It is also natural to associate a digital tree to an input word, by interpreting the pointers of the compression algorithm as connections between parents and children. The tree associated to our example string is shown below. The natural size m of a digital

tree is its number of nodes, which corresponds to the number of insertions made and to the number of phrases in the compressing algorithm. Moreover, the internal path length L_m corresponds to the size n of the input word. Therefore, we have

$$M_n = \max\{m | L_m \leq n\}. \tag{1}$$

If we suppose that the insertions are made randomly (this will be made precise), then L_m is a random variable depending on m and from (1) it follows that the random variables M_n in the compression model and L_m in the digital tree model are linked by the formula

$$\Pr(M_n \geq m) = \Pr(L_m \leq n).$$

Furthermore, the relation (1) is known as the renewal equation and from standard probability theory [1], it follows that if L_m has a normal limit distribution, then so has M_n, with $EM_n = n/(EL_n/n)$ and $\operatorname{Var} M_n = \operatorname{Var} L_n/(EL_n/n)^{3/2}$. In the next section we will show that this is actually the case.

The probabilistic model we choose is the asymmetric Bernoulli model. That is, we choose independently letters of a binary alphabet with respective probabilities p and q, where $p, q > 0$ and $p + q = 1$. In the digital tree model, this induces a random insertion of the phrases, which occur during the parsing. Now the inner path length of a digital tree is the sum of the inner path lengths of its two subtrees plus its size. If we note $H_m(u) = E[u^{L_m}]$, this yields

$$H_m(u) = u^m \sum_{k=0}^{m-1} \binom{m-1}{k} p^k q^{m-1-k} H_k(u) H_{m-1-k}(u).$$

Finally, by setting $h(z, u) = \sum_m H_m(u) z^m/m!$, we get our functional equation:

$$h_z(z, u) = h(puz, u)h(quz, u), \tag{2}$$

with $h(z, 1) = e^z$ and $h(0, u) = 1$ as boundary conditions.

2. Analysis of the functional equation

We would like to transform the equation, so that it takes an additive form. If, for a moment, we forget about the derivative, this can be done by studying the logarithm of h and the equation can be rewritten as

$$\ln h(z, u) = \ln h(puz, u) + \ln h(quz, u). \tag{3}$$

We have to verify that the logarithm of h exists. In this simple case this is straightforward and we have the bound $\log h(z, u) = O(z^{k(u)})$, where

$$(pu)^{k(u)} + (qu)^{k(u)} = 1.$$

We remark that $k(u) = 1 - \log u/h + O(\log^2 u)$, where $h = -p \log p - q \log q$ denotes the entropy of the alphabet.

In the original case, it is less straightforward to show that the logarithm exists and to obtain a bound. In fact, we will do this, when u belongs to a real neighbourhood $\mathcal{U}$ of 1 and when z belongs to a polynomial cone $\mathcal{C}(D, \delta) = \{x + iy | x \geq 0 \wedge |y| \leq Dx^\delta\}$, where $D \geq 0$ and $0 \leq \delta < 1$. To put the equation (1) in a form like (3), which is easier to manipulate, we need a new auxiliary function $f(z, u) = h(z, u)/h_z(z, u)$. We obtain

$$f_z(z, u) = 1 - \left(\frac{pu}{f(puz, u)} + \frac{qu}{f(quz, u)} \right) f(z, u) \tag{4}$$

Now f has expected polynomial order $O(z^{1-k(u)})$, which should make this equation easier to study. In fact, we have

THEOREM 1. *There exist a convex polynomial cone $\mathcal{C}(D,\delta)$ of z and a real neighbourhood $\mathcal{U}$ of $u=1$, such that $\log h(z,u)$ exists, and $\log h(z,u) = O(z^{k(u)})$, uniformly for $(z,u) \in \mathcal{C}(D,\delta)\times\mathcal{U}$. Moreover, all derivatives of $\log h(z,u)$ with respect to u exist and are of order $O(z^{k(u)+\varepsilon})$, for any $\varepsilon > 0$.*

PROOF. The theorem is at the heart of the analysis, but its proof is quite involved [6]. We will content ourselves with giving some of the main ideas. Let $\mathcal{D}_m$ be the closed disk of centre 0 and radius ρ^{-m}, where $\rho = \max\{p,q\} < 1$. Then $\mathcal{D}_0 \subseteq \mathcal{D}_1 \subseteq \cdots$ and the disks satisfy the fundamental property

$$z \in \mathcal{D}_{m+1} - \mathcal{D}_m \Rightarrow puz, quz \in \mathcal{D}_m, \tag{5}$$

for $m \geq 0$. Moreover, as u is real, for each convex cone $\mathcal{C}(D,\delta)$, a similar property is satisfied:

$$z \in \mathcal{C}(D,\delta) \Rightarrow puz, quz \in \mathcal{C}(D,\delta).$$

This will make it possible to use induction over the domains $\mathcal{D}_m \cap \mathcal{C}(D,\delta)$. Then by using (4) we obtain an integral formula for $f(z,u)$, when $z \in (\mathcal{D}_{m+1} - \mathcal{D}_m) \cap \mathcal{C}(D,\delta)$, where the integrand contains $f(puz,u)$ and $f(quz,u)$, with $puz, quz \in \mathcal{D}_m \cap \mathcal{C}(D,\delta)$. Estimations of this integral are used to prove the theorem. □

As a consequence of the theorem we can bound the error term in the expansion of $h(z,e^t)$ by using Taylor's formula with integral rest. We obtain

$$h(z,e^t) = \exp\left(z + X(z)t + V(z)\frac{t^2}{2} + O(t^3 z^{k(e^t)})\right).$$

The mean $X(z)$ and variance $V(z)$ can be computed in the Poisson model (using the Poisson generating function $\tilde{h}(z,e^t) = h(z,e^t)e^{-z}$), by using Mellin transform techniques [4]. This yields

$$\begin{cases} X(z) = \frac{z\log z}{h} + O(z), \\ V(z) = \frac{z\log^2 z}{h^2} + Az\log z + O(z), \end{cases}$$

where $A = 0$ in the case $p = q = 1/2$. We therefore obtain normal asymptotics

$$h(z, e^{t/\sqrt{V(z)}})e^{tX(z)/\sqrt{V(z)}} = \exp\left(\frac{t^2}{2} + O(z^{k(u)-3/2})\right).$$

To complete the analysis, it is necessary to translate these asymptotic expansions back to obtain the limit distribution for L_m. This is done by Cauchy's formula. We have

$$E[u^{L_m}] = \frac{m!}{2i\pi}\oint \frac{h(z,u)dz}{z^{m+1}}.$$

Here the contour is a big circle. It is possible to show [5] that the contribution of the part of the contour outside the polynomial cone is exponentially small, and in view of the discussion in Section 1, Jacquet and Szpankowski obtain the following theorem [4, 6]:

THEOREM 2. *In the asymmetric Bernoulli model, M_n has a normal limit distribution of mean $EM_n \sim nh/\log_2 n$ and variance $\operatorname{Var} M_n \sim Ah^3 n/\log^2 n$. In the case of a symmetric Bernoulli model ($p = q = 1/2$), the variance becomes $\operatorname{Var} M_n \sim (C + \phi(\log_2 n))n/\log_2^3 n$, where ϕ is a periodic function of small amplitude and period 1.*

We remark that in the case of a symmetric Bernoulli model, A vanishes, and the following term in the asymptotic expansion gives rise to the oscillating function in the result. We also remark that the theory can be generalized to an alphabet with more letters; this would give functional equations of the type

$$h_z(z,u) = h(p_1 zu, u) \cdots h(p_\nu zu, u),$$

with $\nu \geq 2, p_1, \ldots, p_\nu > 0$ and $p_1 + \cdots + p_\nu = 1$. However, ν may not be equal to 1, because the fundamental property (5) for the domains $\mathcal{D}_m$ fails in this case.

Bibliography

[1] Billingsley (Patrick). – *Probability and Measure*. – John Wiley & Sons, 1986, 2nd edition.

[2] Flajolet (P.) and Sedgewick (R.). – Digital search trees revisited. *SIAM Journal on Computing*, vol. 15, n° 3, August 1986, pp. 748–767.

[3] Flajolet (Philippe) and Richmond (Bruce). – Generalized digital trees and their difference-differential equations. *Random Structures and Algorithms*, vol. 3, n° 3, 1992, pp. 305–320.

[4] Jacquet (P.) and Szpankowski (W.). – A functional equation often arising in the analysis of algorithms on words. – Unpublished conference version.

[5] Jacquet (Philippe) and Régnier (Mireille). – *Normal limiting distributions for the size and the external path length of tries*. – Research report n° 827, Institut National de Recherche en Informatique et en Automatique, April 1988.

[6] Jacquet (Philippe) and Szpankowski (Wojciech). – *Asymptotic behavior of the Lempel-Ziv parsing scheme and digital search trees*. – Research report n° 2212, Institut National de Recherche en Informatique et en Automatique, March 1994.

[7] Lempel (A.) and Ziv (J.). – On the complexity of finite sequences. *IEEE Transactions on Information Theory*, vol. 22, n° 1, 1976, pp. 75–81.

[8] Ziv (J.). – Compression, test of randomness and estimating the statistical model of individual sequences. In Capocelli (R.) (editor), *Sequences*, pp. 366–373. – Springer Verlag, New York, 1990.

[9] Ziv (J.) and Lempel (A.). – A universal algorithm for sequencial data compression. *IEEE Transactions on Information Theory*, vol. 23, n° 3, 1977, pp. 337–343.