1987

On An Alternative Sum Useful in the Analysis of Some Data Structures

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Report Number:
87-676

ON AN ALTERNATIVE SUM USEFUL IN THE ANALYSIS OF SOME DATA STRUCTURES

Wojciech Szpankowski

CSD-TR-676
April 1987

ON AN ALTERNATIVE SUM USEFUL IN THE ANALYSIS OF SOME DATA STRUCTURES

Wojciech Szpankowski
Department of Computer Science
Purdue University
West Lafayette, IN 47907

Keywords: digital search trees, conflict resolution algorithms, recurrence equations, alternative sums, asymptotic analysis, Mellin transform.

1. MOTIVATIONS

Digital search trees find many applications in computer science and telecommunications. Among others we mention here : an index file in dynamic hashing [3], partial match retrieval of multidimensional data, radix exchange sort, polynomial factorization, simulation, Huffman's algorithm [2] [9], generating an exponentially distributed variate [5]; also in recently developed conflict resolution algorithms for broadcast communication [7] [10] [12] [13], etc. It is well known [4],[9], [14], [15] that some properties (average complexity) of digital search trees can be studied through a solution of some recurrences. To illustrate it, let us assume that n records are stored in a tree, and keys consist of (possibly infinite) sequence of 0's and 1's; the digit 0 occurs with probability p and 1 with probability $q = 1-p$. Then, for example, all moments of the successful search in radix search tries and Patricia tries satisfy the following recurrence [4] [8] [9] [13] [14].

$$x_n = a_n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [x_n + x_{n-k}], \quad n \geq 2 \tag{1}$$

and x_0, x_1 are given. The sequence a_n is called *additive term* and various properties of tries can be modelled by appropriate choice of a_n. For digital search tree [3] [15] (records are stored in internal nodes instead of external nodes as in radix tries and Patricia tries) the equivalent recurrence is

$$x_n = a_n + \sum_{k=0}^{n-1} \binom{n-1}{k} p^k q^{n-1-k} [x_k + x_{n-1-k}] \quad n \geq 2 \tag{2}$$

with x_0 and x_1 given. Some other properties as the average number of internal nodes, the average number of nodes with both sons null and so on, satisfy recurrences (1) and (2), too. Recurrence (1) finds also applications in performance evaluation of the so called stack protocols for conflict resolution algorithms [7] [10] [12].

Another recurrence of similar type, however more sophisticated, appears in the unsuccessful search in a Patricia trie [5] [8] [9] [13] and the analysis of the so called interval-searching algorithms in broadcast communication [13]. Namely:

$$(2^{n+s} - 2)x_n = 2^n a_n + \sum_{k=1}^{n-1} \binom{n}{k} x_k \tag{3}$$

where x_0 , x_1 are given, s is an integer, and a_n is any sequence which models different properties of the trees.

It can be proved that all three recurrences have a common pattern for the solution, namely [4] [12] [13] [14]

$$x_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} f_k \tag{4}$$

where f_k is different for (1)–(3). For example, in [12] it is proved that (1) possesses a solution

$$x_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{\hat{a}_k + k a_1 - a_0}{1 - p^k - q^k} \tag{5}$$

for $x_0 = x_1 = 0$, where $\hat{a}_n$ is the so called inverse relation (or binomial transform) of a_n [11], that is, $\hat{a}_n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} a_k$. The recurrence (3) is more intricate (see [13]), and the so called Bernoulli inverse relation is involved in the solution. For example, for $a_n = q^n$, $0 < q \leq 1$, [13] [15]

$$x_n = \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{B_{k+1}(1-q)}{k+1} \frac{1}{2^{k+s-1} - 1} \tag{6}$$

where $B_k(x)$ is the Bernoulli polynomial [1].

From the average complexity view point, the most interesting situation occurs when n becomes large. We investigate an asymptotic approximation of x_n given by (4) for any f_k. This paper provides a general solution to this problem. The previous solutions have been restricted to very particular cases. For example, for the alternative sum of form (5), with the numerator replaced by a constant, de Bruijn (see [9], 5.2.2) has suggested to develop the denominator into a geometric series, then after a lengthy algebraic manipulation, to obtain a sum with a term $(1 - x/n)^n$ which is next approximated by e^{-x}. Finally, the Mellin transform replaces e^{-x} and the residue theorem is recalled to obtain asymptotic approximation. Our approach is quite different. The final formula is Mellin like, but we do not explicitly use the Mellin transform.

2. MAIN RESULTS

Let us consider an alternative sum

$$S_m(n) \stackrel{\text{def}}{=} \sum_{k=m}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] f_k \tag{7}$$

where f_k is any sequence of numbers. Assume now that f_k has an analytical continuation to a complex function $f(z)$, that is, $f(k) = f_k$. To present our results in a compact form, let us also define for $\alpha, \beta > 0$

$$K_1(\alpha) = \int_{-\infty}^{C} (\alpha/x)^x \, |f(-x \mp i\alpha)| \, dx \qquad C > 0 \tag{8a}$$

$$K_2(\beta) = \int_{-\infty}^{\infty} e^{-|y|} \, |f(\beta - ½ - iy)| \, dy \tag{8b}$$

We prove

Theorem 1. If $f(z)$ is an analytic function left to the line $(½ - m - i\infty,\ ½ - m + i\infty)$, then

$$S_m(n) = \frac{1}{2\pi i} \int_{½ - m - i\infty}^{½ - m + i\infty} \frac{\Gamma(z) f(-z)}{n^{\underline{z}}} \, dz \tag{9}$$

provided that

$$K_1(\alpha) = o(e^{\alpha}) \; ; \qquad K_2(\beta) = o(\beta^{\beta}) \tag{10}$$

where $n^{\underline{z}} \overset{def}{=} \dfrac{\Gamma(n+1)}{\Gamma(n+1-z)}$, $\Gamma(z)$ is the gamma function [6].

Proof. We evaluate the integral in (9) by the Cauchy Theorem . Let us consider a large rectangle $R_{\alpha\beta}$, left to the line of integration, with corners at four points $(\frac{1}{2} - \beta \mp i\alpha,\ C \mp i\infty)$, α, $\beta \geq 0$ and $C = \frac{1}{2} - m$. Then, by the Cauchy Theorem [6] the integral in (9) is equal to the sum of residues in $R_{\alpha\beta}$ minus the values of the integral on the bottom, top and left lines of $R_{\alpha\beta}$, with $\alpha, \beta \to \infty$. We first evaluate these integrals, that is, on the bottom, top and left lines, which are further denoted as I_B, I_T and I_L respectively. Consider first I_B. After simple manipulation one shows that

$$I_B = \int_{-\infty}^{C} \frac{\Gamma(x+i\alpha)\Gamma(n+1-z)}{\Gamma(n+1)} f(-x-i\alpha)dx$$

with $z = x + i\alpha$. But [6]

$$\frac{\Gamma(n+1-z)}{\Gamma(n+1)} = O(n^{-z}) \; ; \quad \Gamma(x+i\alpha) = O(\alpha^{x-\frac{1}{2}} e^{-\pi\alpha/2}) \tag{11}$$

Hence

$$|I_B| = O(e^{-\alpha} \int_{-\infty}^{C} (\alpha/n)^x \, |f(-x-i\alpha)| \, dx)$$

Now note that under condition (10) for $K_1(\alpha)$ $I_B \to 0$ as $\alpha \to \infty$. In similar ways, we prove $I_T \to 0$ for $\alpha \to \infty$. For the integral on the left line, one finds with $\alpha \to \infty$

$$I_L = \int_{-\infty}^{\infty} \Gamma(\tfrac{1}{2} - \beta - iy) \, \frac{\Gamma(n+1-z)}{\Gamma(n+1)} \, f(\beta - \tfrac{1}{2} - iy)dy$$

To evaluate I_L we use (11) together with [6] [9]

$$\Gamma(\tfrac{1}{2} - \beta - iy) = \frac{\Gamma(\tfrac{1}{2} - iy)}{O(\beta!)}$$

Hence

$$|I_L| = O\left(\frac{n^{\beta}}{\beta!} \int_{-\infty}^{\infty} |\Gamma(\tfrac{1}{2} + iy)| \; |f(\beta - \tfrac{1}{2} - iy)| \, dy \right)$$

Using the Stirling approximation for $\beta!$ and (11), we prove that $I_L \to 0$ for $\beta \to \infty$ provided (10) holds for $K_2(\beta)$.

It remains to compute residues left to the line of integration, that is, in $R_{\alpha\beta}$ for $\alpha\rightarrow\infty$, $\beta\rightarrow\infty$. By the assumption that $f(z)$ is analytic in $R_{\alpha,\beta}$, hence by Cauchy Theorem, the integral is equal to the sum of residues at points $z_k = -k$, $k \geq m$ (k is an integer), where the gamma function $\Gamma(z)$ has singularities of value $\frac{(-1)^k}{k!}$ [6]. This implies

$$\frac{1}{2i\pi}\int_{C-i\infty}^{C-i\infty} \frac{\Gamma(z)f(-z)}{n^{\underline{z}}}\,dz = \sum_{k=m}^{\infty} \frac{(-1)^k}{k!} f(-(-k))n^{\underline{k}}$$

But $n^{\underline{k}} = n(n-1)..(n-1+k) = \frac{n!}{(n-k)!}$, hence (9) immediately follows.

□

The function $n^{\underline{z}}$ is hard to analyze, and therefore, a simplification of (9) might be useful. We prove

Theorem 2. Under the assumptions of Theorem 1, the following holds

$$S_m(n) = \frac{1}{2i\pi}\int_{\frac{1}{2}-m-i\infty}^{\frac{1}{2}-m+i\infty} \Gamma(z)f(-z)n^{-z}\,dz + e_n \tag{12}$$

where

$$e_n = O(n^{-1})\frac{1}{2i\pi}\int_{\frac{1}{2}-m-i\infty}^{\frac{1}{2}-m+i\infty} z\Gamma(z)f(-z)n^{-z}\,dz \tag{13}$$

Proof. Eqs. (12) and (13) directly follow from [1]

$$\frac{1}{n^{\underline{z}}} = \frac{\Gamma(n+1-z)}{\Gamma(n+1)} = n^{-z}[1 + zO(n^{-1})]$$

□

To simplify further the computation , it is convenient to consider the following generalization of $S_m(n)$

$$S_{m,r}(n) \stackrel{def}{=} \sum_{k=m}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix}\right] \left\{\begin{matrix} k \\ r \end{matrix}\right\} f(k) \tag{14}$$

where r is an integer. Then

Corollary 1. If hypotheses of Theorem 1 hold, then

$$S_{m,r}(n+r) = (-1)^r \binom{n+r}{r} \frac{1}{2i\pi} \int_{\frac12 - [m-r]^+ - i\infty}^{\frac12 - [m-r]^+ + i\infty} \frac{\Gamma(z) f(r-z)}{n^z}\, dz \tag{15}$$

and

$$S_{m,r}(n) = \frac{(-1)^r}{r!} \frac{1}{2i\pi} \int_{\frac12 - [m-r]^+ - i\infty}^{\frac12 - [m-r]^+ + i\infty} \Gamma(z) f(r-z) n^{r-z}\, dz + e_n \tag{16}$$

with e_n given by (13) with some obvious modifications.

Proof. Note that

$$S_{n,r}(n+r) = \sum_{k=m}^{n+r} (-1)^k \binom{n+r}{k} \binom{k}{r} f(k) = \binom{n+r}{r} (-1)^r \sum_{k=[m-r]^+}^{n} (-1)^k \binom{n}{k} f(k+r).$$

Then, (15) and (16) follow from Theorem 1 and 2.

□

For *asymptotic analysis* of $S_{m,r}(n)$ we use (9) or (15), but this time we integrate along a path which extends far to the *right* from the line of the integration. More precisely, define a large rectangle $R'_{\alpha\beta}$ right to the line $(C - i\infty, C + \infty)$, $C = \frac12 - [m-r]^+$, with corners $(\frac12 + \beta \mp i\alpha \,;\, C \pm i\alpha)$. As before, the integrals on the bottom and top lines of $R'_{\alpha\beta}$ are small for $\alpha \to \infty$ by the same arguments as above. The integral I_R, on the right line can be estimated as

$$|I_R| = O\left[\beta!\, n^{-\beta} \int_{-\infty}^{\infty} e^{-|y|} |f(-\tfrac12 - \beta - iy)|\, dy \right] \tag{17}$$

assuming $\alpha \to \infty$. For fixed but large β, we find $I_R = O(n^{-\beta})$, hence for large n the integral I_R has negligible contribution, and $S_{m,r}(n)$ is equal to the sum of residues right to the line of integration. In the next section, we show how this idea can be applied to evaluate some alternative sums arising in the analysis of algorithms. Finally, let us mention that an alternative approach, using Rice's method, to obtain an asymptotic approximation of the sum (7) is discussed in [4].

3. APPLICATION TO ASYMPTOTIC ANALYSIS

As explained in the first section, a sum of the form

$$S_r(n) = \sum_{k=2}^{r} (-1)^k \binom{n}{k} \binom{k}{r} \frac{1}{1-p^k-q^k} \tag{18}$$

finds applications in the analysis of successful search in radix tries and Patricia tries as well as in conflict resolution algorithms. By (15) we find

$$S_r(n) = \frac{(-1)^r}{r!} \frac{n}{2\pi i} \int_{\frac{1}{2}-[2-r]^+-i\infty}^{\frac{1}{2}-[2-r]^++i\infty} \frac{\Gamma(z) n^{r-1-z}}{1-p^{r-z}-q^{r-z}} dz + e_n \tag{19}$$

where $p + q = 1$. The function under the integral has (right to the line of integration) poles at zeros of the denominator, that is, for $z_k^r = r-1+ix_k$, $k = 0, \pm 1, \pm 2, \ldots,$ with $x_0 = 0$. For $r = 0$ and $r = 1$ the roots $z_0^0 = -1$ and $z_0^1 = 0$ are also poles of the gamma function. It is known that the main contribution to the asymptotic approximation comes from the real roots z_k^r, that is, for $k = 0$. To compute the residues for $k = 0$, we use the following Taylor expressions for $w = z - z_0^r = z + 1 - r$ [12]

$$n^{r-1-z} = 1 - w \ln n + O(w^2) \tag{20a}$$

$$\frac{1}{1-p^{r-z}-q^{r-z}} = -\frac{w^{-1}}{h_1} + \frac{h_2}{2h_1^2} + O(w) \tag{20b}$$

where $h_i = (-1)^i [p \ln^i p + q \ln^i q]$. In addition, we have [6] [9] [12]

- for $r = 0$ $(z_0^0 = -1)$

$$\Gamma(z) = -w^{-1} + (\gamma - 1) + O(w) \tag{21a}$$

- for $r = 1$ $(z_0^1 = 0)$

$$\Gamma(z) = z^{-1} - \gamma + O(z) \tag{21b}$$

- for $r \geq 2$

$$\Gamma(z) = (r-1)! + O(w) \tag{21c}$$

where γ=0.577 is the Euler constant. To find the residue at z_0^r, we multiply (20)–(21) and identify the coefficient of w^{-1}. In a similar way, residues at z_k^r, $k \neq 0$ are obtained. Then after simple algebra

$$S_r(n) = \begin{cases} n\left\{ \dfrac{\ln n + \gamma - \delta_{r,0}}{h_1} + \dfrac{h_2}{2h_1^2} + (-1)^r g_r(n) \right\} + e_n & r=0,1 \\ (-1)^r\, n\left\{ \dfrac{1}{r(r-1)h_1} + g_r(n) \right\} + e_n & r \geq 2 \end{cases} \tag{22}$$

where $g_r(n)$ is the contribution from z_k^r, $k \neq 0$, and it is proved that this is a periodic function with a very small amplitude [3] [8] [9] [12]. For example, if $p=q=0.5$ then [9]

$$g_r(n) = \frac{1}{\ln 2} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \Gamma(r+2\pi ik/\ln 2)\, \exp[-2\pi ik \log_2 n]$$

To complete the analysis, it remains to evaluate e_n given by (13). But we can use exactly the same computation as before noting that the term z in e_n "cancels" the term $\ln n$ in (22), hence the integral is of order $O(n)$, and $e_n = O(1)$. We note also that the proposed evaluation of e_n is much simpler than by traditional techniques (see [9] [12]).

The last example deals with the sum of the form (6). Let for $s \geq 0$

$$R_s(n) = \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{-B_{k+1}(1-q)}{k+1} \frac{1}{2^{k+s}-1} \tag{23}$$

To apply Theorem 1, we need an analytical continuation of $-B_{k+1}(1-q)/(k+1)$. Fortunately, it is known that [1]

$$\zeta(-k, q) = -\frac{B_{k+1}(q)}{k+1}$$

where $\zeta(z, q)$ is the generalized Riemman zeta function [1] [6] [13]. Hence, by (9)

$$R_s(n) = \frac{1}{2i\pi} \int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \frac{\zeta(z, 1-q-\delta_{q,1})\Gamma(z)n^{-z}}{2^{s-z}-1}\, dz + e_n \tag{24}$$

where $\delta_{n,k}$ is the Kronecker delta. Again, we consider the poles of the function under the integral right to the line of integration, that is, zeros of the denominator. Let $z_k = s + 2\pi ik/\ln 2$. The main contribution comes from $z_0 = s$. To compute the residues of z_0, we use the Taylor expansion (21) together with [1] [6] [13]

$$n^{-z} = n^{-s}(1 - w \ln n + O(w^2))$$
$$\frac{1}{2^{s-z} - 1} = \frac{w^{-1}}{\ln 2} - ½ + O(w)$$

where $w = z - s$. For $s = 0$ and $s = 1$, we also need Taylor expansions of the zeta function. But for $s = 0$ [1]

$$\zeta(z, q) = -(½ - q) + z\zeta'(0, q) + O(z^2)$$

while for $s = 1$ with $w = z - 1$

$$\zeta(z, q) = w^{-1} - \psi(q) + O(w)$$

where $\psi(z)$ is the psi function [1]. Finding the coefficient at w^{-1} we compute the residue of the function under the integral. For example, for $s = 1$, we obtain (see [13] for details)

$$R_1(n) = n^{-1} \{lg\ n - ½ + \frac{\gamma}{\ln 2} - \frac{\psi(1 - q - \delta_{q,1})}{\ln 2} + F_1(n) \} + e_n \qquad (25a)$$

and for s=0

$$R_0(n) = (½ + \delta_{q,1} - q)\ (lg\ n - ½ + \gamma/\ln 2) + \zeta'(1 - q + \delta_{q,1})/\ln 2 + F_0(n) + e_n \qquad (25b)$$

where $F_s(n)$ is the contribution from z_k, $k \neq 0$ and one proves [13]

$$F_s(n) = \frac{1}{\ln 2} \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \zeta(s + 2\pi ik/\ln 2)\Gamma(s + 2\pi ik/\ln 2)\exp[-2\pi ik lg\ n] \qquad (26)$$

The most interesting, by some standards, is the evaluation of e_n. Note that in terms of big Oh notation, the contribution to e_n comes from n^{-z} in the integral (24). In the case $s = 0$ and $s = 1$, a term $lg\ n$ has appeared before, but it is ''naturalized'' in e_n by the term z. Hence the integral is $O(n^{-s})$, and $e_n = O(n^{-s-1})$. This fact required a lengthy proof in [9] and [13].

REFERENCES

[1] **Abramowitz, M.**, and **Stegun, L.**, Handbook of mathematical functions, Dover, New York (1964).

[2] **Aho, A.**, **Hopcroft, J.**, and **Ullmann, J.**, Data structures and algorithms, Addison-Wesley (1983).

[3] **Flajolet, Ph.**, On the performance evaluation of extendible hashing and trie searching, *Acta Informatica* 2O (1983), 345–369.

[4] **Flajolet, Ph.** and **Sedgewick, R.**, Digital search trees revisited, *SIAM J. Comput.*, 15 (1986), 748–767.

[5] **Flajolet, Ph.** and **Saheb, N.**, The complexity of generating an exponentially distributed variate, *J. Algorithms*, 7, (1987), 463–488.

[6] **Henrici, P.**, Applied and computational complex analysis, John Wiley & Sons, New York (1977).

[7] **Hofri, M.**, Stack algorithms for collision-detecting channels and their analysis: A limited survey, *Proceedings Intern. Seminar on Modeling and Performance Evaluation Methodology*, Paris (1983), 53–78.

[8] **Kirschenhofer, P.** and **Prodinger, H.**, Some further results on digital trees, *Proceedings of ICALP 86*, Rennes, 1986.

[9] **Knuth, D.**, The art of computer programming, sorting and searching, Addison-Wesley (1973).

[10] **Mathys, P.** and **Flajolet, Ph.**, Q-ary collision resolution algorithms in random-access systems with free and blocked channel access, *IEEE Trans. Information Theory*, vol. IT-31, (1985), 217–243.

[11] **Riordan, J.**, Combinatorial identities, John Wiley & Sons (1968).

[12] **Szpankowski, W.**, On a recurrence equation arising in the analysis of conflict resolution algorithms, *Stochastic Models*, 3, (1987).

[13] **Szpankowski, W.**, Solution of a linear recurrence equation arising in the analysis of some algorithms, *SIAM J. Algebraic and Discrete Methods*, 8, (1987), 233–250.

[14] **Szpankowski, W.**, Average complexity of additive properties for multiway tries: A unified approach, *Proceedings of CAAP'87*, Pisa 1987.

[15] **Szpankowski, W.**, Two problems on the average complexity of digital trees (submitted to a conference).