Compositions and Patricia tries: no fluctuations in the variance![*] [†]

Helmut Prodinger

Dedicated to Hosam Mahmoud on the occasion of his 50^{th} birthday.

Abstract

We prove that the variance of the number of different letters in random words of length n, with letters i and probabilities 2^{-i} attached to them, is $1 + o(1)$. Likewise, the variance of the insertion cost of symmetric Patricia tries of n random data is given by $1 + o(1)$. These two examples disprove popular belief that such quantities must *always* contain fluctuating terms.

1 Introduction

A surprisingly large number of results in *analysis of algorithms* contain *fluctuations*. A typical result might read "The expected number of . . . for large n behaves like $\log_2 n$ + constant + $\delta(\log_2 n)$." Examples include various trie parameters, approximate counting, probabilistic counting, radix exchange sort, leader election, skip lists, adaptive sampling; see the classic books by Flajolet, Knuth, Mahmoud, Sedgewick, Szpankowki [16, 11, 12, 14, 18] for background.

As one can see from Figure 1, $\delta(x)$ has mean zero (the zeroth Fourier coefficient is not there) and very small amplitude. On the other hand, $\delta^2(x)$ is still periodic with period 1, but its mean is *not* zero. Why should we worry about a quantity apparently as small as $\approx 10^{-12}$?

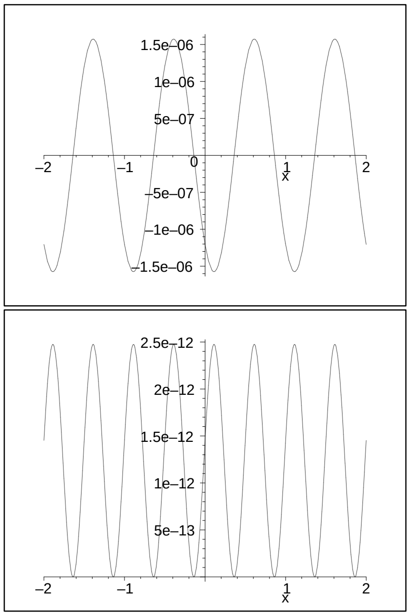

Figure 1: $\delta(x)$ and $\delta^2(x)$

The reason is the *variance* of such parameters, as it naturally contains the term "−expectation^2," and as such also $-\delta^2(x)$. That might not be a sufficient motivation for a casual reader if it were not the case that often *substantial cancellations* occur. In order to identify them, one has to know more about $\delta^2(x)$. If one ignores these terms, one gets wrong results, and the results are not wrong by $\approx 10^{-12}$, but *by an order of growth!* Path length in tries, Patricia tries, and digital search trees [4, 10, 5] are such cases: the variance is in reality of order n only, but ignoring the fluctuations would lead to a (wrong) $\approx n^2$ result.

Size and node level in unbiased (=symmetric) tries exhibit concentration of distribution but proving this requires nontrivial modular form identities, as described for instance in [8]. This in turn has impact on the stability of certain communication protocols—in particular, the tree protocol of Capetanakis–Tsybakov–Mikhailov whose status remained partly unsettled for a few years: see the papers by Berger, Gelenbe and Massey in [13] and the special issue [15].

Now, questions like that occurred in several writings of this author (together with various coauthors), as can be seen from the references. The techniques are extremely interesting, as one has to dig deep into classical analysis. So far, it seems that the *calculus of residues*, as used in the sequel, is the most versatile approach in this context. Another approach is to use (modular) identities due to Dedekind, Ramanujan, Jacobi and others (which can often be proved by Mellin transform techniques); however, often they do not quite *fit.* The residue calculus approach directly addresses the

[*] Supported by NRF Grant 2053748.

[†] The John Knopfmacher Centre for Applicable Analysis and Number Theory, School of Mathematics, University of the Witwatersrand, Private Bag 3, Wits, 2050 Johannesburg, South Africa, `helmut@maths.wits.ac.za`

formula that is ultimately needed.

Many such considerations have been performed about 10–15 years ago, but a new surprise showed up in August 2003: There are two examples, where the variance has *no fluctuation at all* (at least not in the leading term). This does not seem to be intuitive by any means, so we must rely on some analysis to exhibit that phenomenon. The present paper is devoted to just that.

The two sections that follow prove the following theorem.

THEOREM 1.1. *1. Consider words $x_1 \dots x_n$, where the letters follow (independent) geometric random variables X with $\mathbb{P}\{X = i\} = 2^{-i}$ for $i = 1, 2, 3, \dots$. Then the variance of the parameter "number of different letters in a random word of length n" is $1 + o(1)$.*

2. The variance of the insertion cost of a random Patricia trie constructed from n random data is $1 + o(1)$.

The methods that are presented here also allow one to *simplify* the Fourier coefficients of the fluctuations in the variance even in such cases where the periodic function persists. After all, the ultimate simplification is to show that the Fourier coefficients are zero in the two cases on which this paper concentrates.

2 Words and Compositions

In a forthcoming paper [1], words $x_1 \dots x_n$ are considered, where the letters follow (independent) geometric random variables X with $\mathbb{P}\{X = i\} = pq^{i-1}$ for $i = 1, 2, 3, \dots$ and $p + q = 1$. The parameter of interest is the number of different letters in a random word of length n which appear at least b times. Paweł Hitczenko, who visited us in August 2003, reported that he and Guy Louchard [2] had considered the special case $p = q = \frac{1}{2}$ and $b = 1$ in the context of random *compositions*. The variance of the number of part sizes is given by $1 + o(1)$. Our general analysis however predicted a result of the form $\log_Q 2 + \delta_V(\log_Q n) + o(1)$, with $Q = 1/q$ and a periodic function $\delta_V(x)$ of period one. Such oscillations are quite common in *analysis of algorithms* and *enumerative combinatorics*; see e. g., the books [16, 18]. We were both right, and, indeed, in the special case, the periodic function cancels out! This will be demonstrated in the sequel.

After this surprise, I looked for other examples from the past and periodic oscillations in the variance that might actually *not be there*, and I found the instance of (symmetric) *Patricia tries*, which I will treat in the following section.[1]

For interest, the variance is computed from the exponential generating function

$$\Psi(z, u) = \prod_{i \ge 1} \bigl(1 + u(e^{z/2^i} - 1)\bigr)$$

as

$$n![z^n] \frac{\partial^2}{\partial u^2} \Psi(z, u)\Big|_{u=1} + n![z^n] \frac{\partial}{\partial u} \Psi(z, u)\Big|_{u=1} - \Bigl(n![z^n] \frac{\partial}{\partial u} \Psi(z, u)\Big|_{u=1}\Bigr)^2.$$

The periodic function that our analysis exploits, is given as a Fourier series,

$$\delta_V(x) = \sum_{k \ne 0} a_k e^{2\pi i k x},$$

with

$$a_k = \frac{2}{L}\Gamma(-\chi_k)\Bigl[\frac{\psi(-\chi_k) + \gamma}{L} + g(\chi_k)\Bigr] - \frac{1}{L^2} \sum_{j \ne 0, \ne k} \Gamma(-\chi_j)\Gamma(-\chi_{k-j}), \quad k \ne 0;$$

here and in the sequel, we will use the abbreviations $L = \log 2$ and $\chi_k = \frac{2\pi i k}{L}$. Not surprisingly, the sum term originates from the square of a periodic function which was contained in the asymptotic expansion of the *expectation*. The function $g(x)$ is defined by

$$g(x) = \sum_{l \ge 1} \binom{x}{l} \frac{1}{2^l - 1}$$

and $\psi(x)$ is the logarithmic derivative of the Gamma function. Our goal is to show that $a_k = 0$ for all $k \in \mathbb{Z}$, $k \ne 0$.

Let us rewrite the formula for a_k, using not more than the formula $\Gamma(-x)x(x-1)\dots(x-j+1) = (-1)^j \Gamma(-x+j)$:

$$\begin{aligned} L^2 a_k &= 2\Gamma(-\chi_k)\bigl(\psi(-\chi_k) + \gamma\bigr) \\ &\quad + 2L \sum_{l \ge 1} \frac{(-1)^l}{l!(2^l - 1)} \Gamma(l - \chi_k) \\ &\quad - \sum_{j \ne 0, \, \ne k} \Gamma(-\chi_j)\Gamma(-\chi_k + \chi_j). \end{aligned}$$

[1] It is not likely that significantly different examples can be found.

The technique to rewrite the second sum (and similar ones) accordingly was already presented in earlier publications; let me just cite [9]: One considers

$$F(z) = L\frac{\Gamma(z)\Gamma(-\chi_k - z)}{e^{Lz} - 1}$$

and its integral

$$I_1 = \frac{1}{2\pi i}\int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} F(z)dz.$$

The choice of this function is driven by the fact that the denominator has simple poles at $z = \chi_j$ for all $j \in \mathbb{Z}$, and that the numerator produces the "right" residues.

The line of integration will be shifted to $\Re z = -\frac{1}{2}$. There are poles at $z = -\chi_j$, for each $j \in \mathbb{Z}$. Taking them into account, one gets

$$\begin{aligned} I_1 &= \frac{1}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} F(z)dz \\ &+ \sum_{j\neq 0,\,\neq k} \Gamma(-\chi_j)\Gamma(-\chi_k + \chi_j) \\ &- 2\Gamma(-\chi_k)\big(\psi(-\chi_k) + \gamma\big). \end{aligned}$$

Now one writes

$$\frac{1}{e^z - 1} = -1 - \frac{1}{e^{-z} - 1}$$

and gets

$$\begin{aligned} I_1 = &-\frac{L}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \Gamma(z)\Gamma(-\chi_k - z)dz \\ &- \frac{1}{2\pi i}\int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} L\frac{\Gamma(-z)\Gamma(-\chi_k + z)}{e^{Lz} - 1}dz \\ &+ \sum_{j\neq 0,\,\neq k} \Gamma(-\chi_j)\Gamma(-\chi_k + \chi_j) \\ &- 2\Gamma(-\chi_k)\big(\psi(-\chi_k) + \gamma\big). \end{aligned}$$

The simple change of variable $z := z + \chi_k$ produces the integral I_1 again, and one finds

$$\begin{aligned} 2I_1 = &-LI_2 + \sum_{j\neq 0,\,\neq k} \Gamma(-\chi_j)\Gamma(-\chi_k + \chi_j) \\ &- 2\Gamma(-\chi_k)\big(\psi(-\chi_k) + \gamma\big). \end{aligned}$$

What remains is the evaluation of the integral

$$\begin{aligned} I_2 &= \frac{1}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \Gamma(z)\Gamma(-\chi_k - z)dz \\ &= \frac{1}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \Gamma(-\chi_k + z)\Gamma(-z)dz. \end{aligned}$$

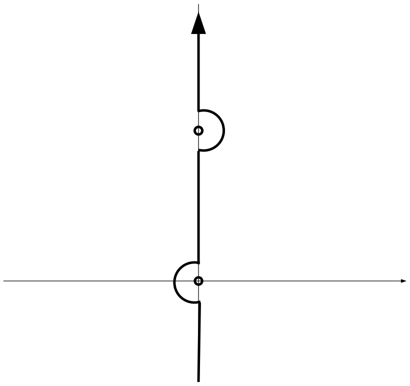

Figure 2: The path of integration.

This integral can be evaluated using the technique of Barnes, as explained in [19, p. 286ff]. The line of integration (see Figure 2) must be shifted to $\Re z = 0$, with the provision that the singularities of $\Gamma(-\chi_k + z)$, i. e., $z = \chi_k$, must lie on the left of the path, and the singularities of $\Gamma(-z)$, i. e., $z = 0$, must lie on the right of the path. The first thing can be achieved by subtracting the residue at $z = \chi_k$, which leads to a term $\Gamma(-\chi_k)$; for the second thing, nothing must be done. So,

$$I_2 = -\Gamma(-\chi_k) + \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \Gamma(-\chi_k + z)\Gamma(-z)dz.$$

Then,

$$\begin{aligned} &\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \Gamma(-\chi_k + z)\Gamma(-z)dz \\ &\quad = \sum_{l\geq 0} \frac{\Gamma(-\chi_k + l)}{l!}(-1)^l = \Gamma(-\chi_k)2^{\chi_k} = \Gamma(-\chi_k). \end{aligned}$$

Altogether, we have seen that $I_2 = 0$. We note that the integral is the sum of the (negative) residues right to the line $\Re z = -\frac{1}{2}$:

$$I_2 = -\Gamma(-\chi_k) + \sum_{l\geq 0} \frac{\Gamma(-\chi_k + l)}{l!}(-1)^l.$$

On the other hand, integral

$$\begin{aligned} I_1 &= \frac{1}{2\pi i}\int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} L\frac{\Gamma(z)\Gamma(-\chi_k - z)}{e^{Lz} - 1}dz \\ &= \frac{1}{2\pi i}\int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} L\frac{\Gamma(-\chi_k + z)\Gamma(-z)}{e^{Lz} - 1}dz \end{aligned}$$

can simply be evaluated by shifting the line of integration to the right, and collecting (negative) residues at l

for $l = 1, 2 \dots$. The result is

$$I_1 = L \sum_{l \ge 1} \frac{(-1)^l \Gamma(-\chi_k + l)}{l!(2^l - 1)}.$$

Putting the two different evaluations together, one sees

$$\begin{aligned} 2I_1 &= 2L \sum_{l \ge 1} \frac{(-1)^l \Gamma(-\chi_k + l)}{l!(2^l - 1)} \\ &= \sum_{j \ne 0,\, \ne k} \Gamma(-\chi_j)\Gamma(-\chi_k + \chi_j) \\ &\quad - 2\Gamma(-\chi_k)\big(\psi(-\chi_k) + \gamma\big), \end{aligned}$$

and this is the identity we wanted to prove.

3 Patricia tries

The variance of the insertion cost of a random Patricia trie constructed from n random data was computed in [7] as

$$\frac{\pi^2}{6L^2} + \frac{1}{12} - \frac{2}{L} \sum_{l \ge 1} \frac{(-1)^{l-1}}{l(2^l - 1)} + \sigma_V(\log_2 n) + o(1),$$

with[2]

$$\begin{aligned} \sigma_V(x) &= \frac{2}{L^2} \sum_{k \ne 0} \Gamma(-1 - \chi_k) e^{2\pi i k x} \\ &- \frac{2}{L^2} \sum_{k \ne 0} (1 + \chi_k)\Gamma(-1 - \chi_k)\big(\psi(-1 - \chi_k) + \gamma\big) e^{2\pi i k x} \\ &- \frac{2}{L} \sum_{k \ne 0} (1 + \chi_k)\Gamma(-1 - \chi_k) g(\chi_k) e^{2\pi i k x} - \big(\delta_E(x)\big)^2, \end{aligned}$$

with the same function $g(x)$ as before, and

$$\delta_E(x) = \frac{1}{L} \sum_{k \ne 0} (1 + \chi_k)\Gamma(-1 - \chi_k) e^{2\pi i k x}.$$

For interest, the variance is computed from the recursion

$$H_n(u) = 2^{1-n} u \sum_{k=0}^{n} \binom{n}{k} H_k(u) - 2^{1-n}(u - 1) H_n(u),$$

$n \ge 2$, $H_0(u) = H_1(u) = 1$, via

$$\frac{H''(1)}{n} + \frac{H'(1)}{n} - \Big(\frac{H'(1)}{n}\Big)^2.$$

These results can be rewritten, using $\Gamma(x + 1) = x\Gamma(x)$ and $\psi(x + 1) = \psi(x) + \frac{1}{x}$:

$$\begin{aligned} \sigma_V(x) &= \frac{2}{L^2} \sum_{k \ne 0} \Gamma(-\chi_k)\big(\psi(-\chi_k) + \gamma\big) e^{2\pi i k x} \\ &\quad + \frac{2}{L} \sum_{k \ne 0} \Gamma(-\chi_k) g(\chi_k) e^{2\pi i k x} - \big(\delta_E(x)\big)^2, \end{aligned}$$

and

$$\delta_E(x) = -\frac{1}{L} \sum_{k \ne 0} \Gamma(-\chi_k) e^{2\pi i k x}.$$

Szpankowski[3] obtained related results in [17].

Now if we look at the coefficient of $e^{2\pi i k x}$, for $k \ne 0$, we find

$$\begin{aligned} &\frac{2}{L^2}\Gamma(-\chi_k)\big(\psi(-\chi_k) + \gamma\big) + \frac{2}{L}\Gamma(-\chi_k) g(\chi_k) \\ &\qquad - \frac{1}{L^2} \sum_{j \ne 0, \ne k} \Gamma(-\chi_j)\Gamma(-\chi_{k-j}). \end{aligned}$$

According to the analysis in the previous section, these coefficients are all equal to zero. Let us finally consider the constant term (for $k = 0$):

$$\begin{aligned} -[\delta_E^2(x)]_0 &= -\frac{1}{L^2} \sum_{k \ne 0} \Gamma(-\chi_k)\Gamma(\chi_k) \\ &= -\frac{\pi^2}{6L^2} - \frac{1}{12} + 1 + \frac{2}{L} \sum_{l \ge 1} \frac{(-1)^{l-1}}{l(2^l - 1)}. \end{aligned}$$

This result appeared in several papers, e. g., in [8]. Let us sketch how one can get it: Consider the function

$$F(z) = \frac{L}{e^{Lz} - 1} \Gamma(-z)\Gamma(z)$$

and its integral

$$I_1 = \frac{1}{2\pi i} \int_{\frac12 - i\infty}^{\frac12 + i\infty} F(z) dz.$$

The line of integration will be shifted to $\Re z = -\frac{1}{2}$. There are poles at $z = \chi_k$, for each $k \in \mathbb{Z}$, and so

$$\begin{aligned} I_1 &= -\frac{\pi^2}{6} - \frac{L^2}{12} + \sum_{k \ne 0} \Gamma(-\chi_k)\Gamma(\chi_k) \\ &\quad + \frac{1}{2\pi i} \int_{-\frac12 - i\infty}^{-\frac12 + i\infty} F(z) dz. \end{aligned}$$

[2] There is a typo; the paper [6] contains the correct version. This paper discusses M-ary Patricia tries, and only for the instance $M = 2$ (binary Patricia tries) do the cancellations occur.

[3] In the early years, we worked on such things independently, not long after that, we became coauthors and friends. For instance, we considered the path length in Patricia tries in [10]. Sure enough, cancellation phenomena showed up, but the variance still contains a periodic fluctuation.

As before,

$$2I_1 = -\frac{\pi^2}{6} - \frac{L^2}{12} + \sum_{k\neq 0} \Gamma(-\chi_k)\Gamma(\chi_k) - LI_2$$

with

$$\begin{aligned} I_2 &= \frac{1}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \Gamma(-z)\Gamma(z)dz \\ &= -\frac{1}{2\pi i}\int_{-\frac{1}{2}-i\infty}^{-\frac{1}{2}+i\infty} \frac{\pi}{z\sin(\pi z)}dz. \end{aligned}$$

This integral is the sum of the (negative) residues right to the line $\Re z = -\frac{1}{2}$, i. e.,

$$I_2 = \sum_{l\geq 1} \frac{(-1)^l}{l} = -L.$$

On the other hand, I_1 can be computed as the sum of the (negative) residues right to the line $\Re z = \frac{1}{2}$, viz.

$$I_1 = -\sum_{l\geq 1} \frac{(-1)^l}{l(2^l-1)};$$

the two different evaluations give the identity.

Putting everything together, we find that the variance of the insertion cost of a Patricia tree constructed from n random data is just $1 + o(1)$.

Patricia tries surprised this author in 1986 when it turned out that the constant

$$\frac{\pi^2}{6L^2} + \frac{1}{12} - \frac{2}{L}\sum_{l\geq 1} \frac{(-1)^{l-1}}{l(2^l-1)}$$

is just $1.000000000000\mathit{1237}\ldots$. That was nicely explained by Johannes Schoißengeier [3].

Now, thanks to Paweł Hitczenko, who made me think about Patricia tries (and related material) again, they offered a new surprise in 2003.

Acknowledgment. I thank Margaret Archibald for the critical reading of an earlier draft.

References

[1] M. Archibald, A. Knopfmacher, and H. Prodinger. The number of distinct values in a geometrically distributed sample. *In preparation*, 2003.

[2] P. Hitczenko and G. Louchard. Distinctness of compositions of an integer: A probabilistic analysis. *Random Structures & Algorithms*, 19:407–437, 2001.

[3] P. Kirschenhofer, H. Prodinger, and J. Schoißengeier. Zur Auswertung gewisser numerischer Reihen mit Hilfe modularer Funktionen. In E. Hlawka, editor, *Zahlentheoretische Analysis II*, volume 1262 of *Lecture Notes in Mathematics*, pages 108–110, 1987.

[4] P. Kirschenhofer, H. Prodinger, and W. Szpankowski. On the variance of the external path length in a symmetric digital trie. *Discrete Applied Mathematics*, 25:129–143, 1989.

[5] P. Kirschenhofer, H. Prodinger, and W. Szpankowski. Digital search trees again revisited: The internal path length perspective. *SIAM Journal on Computing*, 23:598–616, 1994.

[6] P. Kirschenhofer and H. Prodinger. Asymptotische Untersuchungen über charakteristische Parameter von Suchbäumen. In E. Hlawka, editor, *Zahlentheoretische Analysis II*, volume 1262 of *Lecture Notes in Mathematics*, pages 93–107, 1987.

[7] P. Kirschenhofer and H. Prodinger. Further results on digital search trees. *Theoret. Comput. Sci.*, 58:143–154, 1988.

[8] P. Kirschenhofer and H. Prodinger. On some applications of formulæ of Ramanujan in the analysis of algorithms. *Mathematika*, 38:14–33, 1991.

[9] P. Kirschenhofer and H. Prodinger. A result in order statistics related to probabilistic counting. *Computing*, 51:15–27, 1993.

[10] P. Kirschenhofer, H. Prodinger, and W. Szpankowski. On the balance property of Patricia tries: external path length viewpoint. *Theoret. Comput. Sci.*, 68:1–17, 1989.

[11] D. E. Knuth. *The Art of Computer Programming*, volume 1: Fundamental Algorithms. Addison-Wesley, 1973. Third edition, 1997.

[12] D. E. Knuth. *The Art of Computer Programming*, volume 3: Sorting and Searching. Addison-Wesley, 1973. Second edition, 1998.

[13] G. Longo, editor. *Multi–User Communication Systems*, volume 265 of *CISM Courses and Lecture Notes.* Springer Verlag, 1981.

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

[15] J. Massey, editor. *Random Access Communication.* I.E.E.E Press, 1985.

[16] R. Sedgewick and P. Flajolet. *An Introduction to the Analysis of Algorithms.* Addison-Wesley, 1996.

[17] W. Szpankowski. Patricia tries again revisited. *J. Assoc. Comput. Mach.*, 37:691–711, 1990.

[18] W. Szpankowski. *Average case analysis of algorithms on sequences.* Wiley-Interscience, New York, 2001.

[19] E. T. Whittaker and G. N. Watson. *A Course of Modern Analysis.* Cambridge University Press, fourth edition, 1927. Reprinted 1973.