Theoretical Computer Science 527 (2014) 1–36

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

An analytic approach to the asymptotic variance of trie statistics and related structures

Michael Fuchs [a,*], Hsien-Kuei Hwang [b], Vytas Zacharovas [c]

[a] Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan
[b] Institute of Statistical Science, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
[c] Dept. Mathematics & Informatics, Vilnius University, Lithuania

A R T I C L E   I N F O

Article history:
Received 7 August 2013
Received in revised form 7 January 2014
Accepted 21 January 2014
Communicated by H. Prodinger

This paper is dedicated to the memory of Philippe Flajolet, who pioneered the asymptotic study of binomial splitting processes

Keywords:
Digital trees
Binomial splitting process
Mellin transform
Variance
Periodic fluctuations
Contention resolution algorithms

A B S T R A C T

We develop analytic tools for the asymptotics of general trie statistics, which are particularly advantageous for clarifying the asymptotic variance. Many concrete examples are discussed for which new Fourier expansions are given. The tools are also useful for other splitting processes with an underlying binomial distribution. We specially highlight Philippe Flajolet's contribution in the analysis of these random structures.

1. Introduction

Coin-flipping is one of the simplest ways of resolving a conflict, deciding between two alternatives, and generating random phenomena. It has been widely adopted in many daily-life situations and scientific disciplines. There exists even a term "flippism." The curiosity of understanding the randomness behind throwing coins or dices was one of the motivating origins of early probability theory, culminating in the classical book "*Ars Conjectandi*" by Jacob Bernoulli, which was published exactly three hundred years ago in 1713 (many years after its completion; see [92,101]). When flipped successively, one naturally encounters the binomial distribution, which is pervasive in many splitting processes and branching algorithms whose analysis was largely developed and clarified through Philippe Flajolet's works, notably in the early 1980s, an important period marking the upsurge of the use of complex-analytic tools in the Analysis of Algorithms.

Technical content of this paper. This paper is a sequel to [55] and we will develop an analytic approach that is especially useful for characterizing the asymptotics of the mean and the variance of additive statistics of random tries under the Bernoulli model; such statistics can often be computed recursively by

$$X_n \stackrel{d}{=} X_{I_n} + X^*_{n-I_n} + T_n, \tag{1}$$

* Corresponding author. Tel.: +886 3 5712121x56461; fax: +886 3 5724679.
E-mail addresses: mfuchs@math.nctu.edu.tw (M. Fuchs), hkwhang@stat.sinica.edu.tw (H.-K. Hwang), vytas.zacharovas@mif.vu.lt (V. Zacharovas).

0304-3975/$ – see front matter
http://dx.doi.org/10.1016/j.tcs.2014.01.024

with suitable initial conditions, where T_n is known, X_n^* is an independent copy of X_n and I_n is the binomial distribution with mean pn, $0 < p < 1$.

Many asymptotic approximations are known in the literature for the variance of X_n, which has in many cases of interest the pattern

$$\frac{\mathbb{V}(X_n)}{n} = c\log n + c' + \left\{ \begin{array}{ll} P(\varpi \log n), & \text{if } \frac{\log p}{\log q} \in \mathbb{Q} \\ 0, & \text{if } \frac{\log p}{\log q} \notin \mathbb{Q} \end{array} \right\} + o(1), \tag{2}$$

where c may be zero, ϖ depends on the ratio $\frac{\log p}{\log q}$ and $P(x) = P(x+1)$ is a bounded periodic function. However, known expressions in the literature for the periodic function P are rare due to the complexity of the problem, and are often either less transparent, or less explicit, or too messy to be stated. In many situations they are given in the form of one periodic function minus the square of the other. The approach developed here, in contrast, provides not only a systematic derivation of the asymptotic approximation (2) but also a simpler, explicit, independent expression for P, notably in the symmetric case ($p = q$). Further refinement of the $o(1)$-term lies outside the scope of this paper and can be dealt with by the approach developed by Flajolet et al. in [34].

Binomial splitting processes. In general, the simple splitting idea behind the recursive random variable (1) (0 going to the left and 1 going to the right) has also been widely adopted in many different modeling processes, which, for simplicity, will be vaguely referred to as "*binomial splitting processes*" (BSPs), where binomial distribution and some of its extensions are naturally involved in the analysis; see Fig. 1 for concrete examples of BSPs that are related to our analysis here. For convenience of presentation, we roughly group these structures in four categories: Data Structures, Algorithms, Collision Resolution Protocols, and Random Models.

To see how such BSPs in different areas can be analyzed, we start from the recurrence ($q = 1 - p$)

$$a_n = \sum_{0 \leqslant k \leqslant n} \pi_{n,k}(a_k + a_{n-k}) + b_n, \quad \text{where } \pi_{n,k} := \binom{n}{k} p^k q^{n-k}, \tag{3}$$

which results, for example, from (1) by taking expectation. Here the "toll-function" b_n may itself involve a_j ($j = 0, 1, \ldots$) but with multipliers that are exponentially small.

From an analytic point of view, the trie recurrence (3) translates for the Poisson generating function

$$\tilde{f}(z) := e^{-z} \sum_{n \geqslant 0} \frac{a_n}{n!} z^n, \tag{4}$$

into the *trie functional equation*

$$\tilde{f}(z) = \tilde{f}(pz) + \tilde{f}(qz) + \tilde{g}(z), \tag{5}$$

with suitable initial conditions. Such a functional equation is a special case of the more general pattern

$$\sum_{0 \leqslant j \leqslant b} \binom{b}{j} \tilde{f}^{(j)}(z) = \alpha \tilde{f}(pz + \lambda) + \beta \tilde{f}(qz + \lambda) + \tilde{g}(z), \tag{6}$$

where $b = 0, 1, \ldots,$ and $\tilde{g}$ itself may involve $\tilde{f}$ and its derivatives $\tilde{f}^{(j)}$ but with exponentially small factors. When $b = 0$, one has a pure functional equation,

$$\tilde{f}(z) = \alpha \tilde{f}(pz + \lambda) + \beta \tilde{f}(qz + \lambda) + \tilde{g}(z), \tag{7}$$

while when $b \geqslant 1$, one has a differential–functional equation.

It turns out that Eq. (6) covers almost all cases we collected (a few hundred of publications) in the analysis of BSPs the majority of which correspond to the case $b = \lambda = 0$. The cases when $b = 0$ and $\lambda > 0$ are thoroughly treated in [21,22,57,81], and the cases when $b \geqslant 1$ are discussed in detail in [55] (see also the references cited there). We focus on $b = \lambda = 0$ in this paper. Since the literature abounds with Eq. (5) or the corresponding recurrence (3), we content ourselves with listing below some references that are either standard, representative or more closely connected to our study here. See also [17,19,41] for some non-random contexts where (5) appeared.

Data structures. Tries: [74,78,104]; PATRICIA tries: [15,74,70]; Quadtries and k-d tries: [32,43]; Hashing: [20,24,39,83]; Suffix trees: [61,104].

Algorithms. Radix-exchange sort: [74]; Bucket selection and bucket sort: [8,77]; Probabilistic counting schemes: [25,27,28,30,90]; Polynomial factorization: [39]; Exponential variate generation: [35]; Group testing: [47]; Random generation: [31,97].