Theoretical Computer Science 93 (1992) 201–225
Elsevier

Dynamic algorithms in D.E. Knuth's model: a probabilistic analysis*,**

G. Louchard

Laboratoire d'Informatique Théorique Université Libre de Bruxelles, B-1050 Brussels, Belgium

B. Randrianarimanana and R. Schott

C.R.I.N., Université Nancy 1, 54506 Vandoeuvre-lès-Nancy, France

Communicated by J. Diaz
Received April 1989
Revised January 1990

Abstract

Louchard, G., B. Randrianarimanana and R. Schott, Dynamic algorithms in D.E. Knuth's model: a probabilistic analysis, Theoretical Computer Science 93 (1992) 201–225.

By dynamic algorithms we mean algorithms that operate on dynamically varying data structures (dictionaries, priority queues, linear lists) subject to insertions I, deletions D, positive (negative) queries Q^+ (Q^-). Let us remember that *dictionaries* are implementable by unsorted or sorted lists, binary search trees, *priority queues* by sorted lists, binary search trees, binary tournaments, pagodas, binomial queues and *linear lists* by sorted or unsorted lists, etc. At this point the following question is very natural in computer science: for a given data structure, which representation is the most efficient? In comparing the space or time costs of two data organizations A and B for the same operations, we cannot merely compare the costs of individual operations for data of given sizes: A may be better than B on some data, and vice versa on others. A reasonable way to measure the efficiency of a data organization is to consider sequences of operations on the structure. Françon (1978, 1979) Knuth (1977) discovered that the number of possibilities for the ith insertion or negative query is equal to i, but that for deletions and positive queries this number depends on the size of the data structure. Answering the questions raised by Françon and Knuth is the main object of this paper. More precisely, we show

- how to obtain limiting processes;
- how to compute explicitly the average costs;
- how to obtain variance estimates;
- that the costs converge as $n \to \infty$ to random variables, either *Gaussian* or depending on *Brownian excursion functionals* (the limiting distributions are, therefore, completely described).

  To our knowledge such a complete analysis has never been done before for dynamic algorithms in Knuth's model.

* This research was (partially) supported by the ESPIRIT 2 Basic Research Actions Program of the EC under contract number 3075 (Alcom project) and by the P.R.C. "Mathématiques et Informatique".

** A preliminary version of our results was included in the proceedings of ICALP'89 conference [20].

1. Introduction

The difficulty of analyzing dynamic algorithms, even if the universe of keys is finite, has been explained by Jonassen and Knuth in [10], where random insertions and deletions are performed on trees whose size never exceeds three. It was shown by Françon [6, 7] and Flajolet et al. [4, 5] that several list and tree organizations can be analyzed in a dynamic context. Integrated costs for these dynamic structures were defined as averages of costs taken over the set of all possible evolutions of the structure, considered up to order isomorphism. Using a method of continued fractions and orthogonal polynomials Flajolet et al. obtained explicit expressions for the expected costs and in some cases for the variances but with Markovian model, which is briefly described in Section 2. The asymptotic distributions were obtained by Louchard [19] with a probabilistic analysis. Taking account of the remarks made by Françon and Knuth, a more natural model has been introduced in [8, 9, 21]: the number of possibilities for the ith insertion or negative query is equal to i, but if after some operations the structure contains k records, the number of possibilities for a deletion or positive query is a linear function of k. Since we have to work with two indices (i and k), the analysis of dynamic algorithms is more difficult in this model. The purpose of the present paper is to give the limiting distributions of cost functions of the linear lists, priority queues and dictionaries. This paper is organized as follows. Section 2 describes the two models. Section 3 provides the set of necessary definitions for dynamic data structures. Section 4.1 is devoted to the linear lists. In Section 4.2 we analyze priority queues and in Section 4.3 the dictionaries. Section 5 provides the main steps of the proofs. Section 6 concludes the paper. An appendix gives some more technical proofs for the dictionary model.

2. The two models

Knuth [13] considers the following operations on a data structure containing k keys (or numbers):

(i) D_r, standing for random deletion, in the sense that if k keys are present, each is chosen for deletion with probability $1/k$;

(ii) D_q, standing for priority queue deletion, i.e. deletion of the smallest key;

(iii) I_0, standing for insertion of a random number by order, in the sense that the new number is equally likely to fall into any of the $k+1$ intervals defined by the k numbers still present as keys after previous insertions and deletions; this is to be independent of the history by which these k numbers were actually obtained;

(iv) I, standing for insertion of a random real number from the uniform distribution in the interval [0, 1]. Each random number inserted is assumed to have the same distribution, and it is to be independent of all previously inserted numbers. Thus, if we look at n such random numbers, the $n!$ possible orderings (of these numbers) are equally likely, and the particular distribution involved has no effect on the behaviour of the data organization (i.e. the class of data structures together with the associated algorithms for operating on these structures).

Knott [11] has shown that I_0 is a concept different from I (see also [4, Section 6.2.2]); this result has stimulated further research, notably [8, 9, 21] and the present work. In this paper, consideration of only the I (I_0) kind of insertion is called *Knuth's* (K) *model* (*Markovian* (M) *model*). As an example, consider a priority queue implemented as a sorted list. Let us analyze the following set of operations (key values are indicated): I(2.5), I(7.4), Dmin, I(1.0), Dmin, I(3.5), Dmin, Dmin.

In the *Markovian model* the first key has order 0, the second key has order 1, the suppression is related to order 0, etc.

In *Knuth's model* the first three operations are identical to the M model. We then have three intervals defined by the first two keys: $(-\infty, 2.5)$, $(2.5, 7.4)$, $(7.4, +\infty)$. The fourth operation I(1.0) is done in the first interval (order 0) among 3 possible intervals. The sixth operation I(3.5) is done in the third interval (order 3) among the 4 possible intervals defined by $1.0<2.5<7.4$, etc.

The Markovian model has been introduced and studied by combinatorial methods in [6, 7]; Flajolet [3] has shown how the theory of continued fractions and orthogonal polynomials remarkably fits this model; further developments appear in [4, 5]; distributions of costs and average costs have been calculated for some sequences of operations for various data types, including priority queues and linear lists. The asymptotic distributions of the costs functions have been obtained by Louchard [19].

The following questions were raised in [7]: how to compute the corresponding costs in Knuth's model and are the costs sensitive to the model? The first answers for linear lists and priority queues were given in [8, 22], after reducing the calculations in Knuth's model to calculations in the Markovian model (we will denote by $\equiv$ the combinatorial equivalence given in [8]).

In [9] an algebraic method has been developed which permits to reproduce all the results of [8] and to treat the dictionary case. In this paper we present a probabilistic analysis of linear lists, priority queues and dictionaries in this model and characterize the corresponding limiting distributions. The first step is to express the problem in a combinatorial way. Following [10], let us consider the sequence of operations III(DI)*, the initial data structure being empty; let $x<y<z$ be the three keys inserted during the sequence III; let us consider a linear list, i.e. x or y or z is deleted with equal probability; let w be the key inserted by the fourth I of this sequence; then all four cases $w<x<y<z$, $x<w<y<z$, $x<y<w<z$, $x<y<z<w$ do occur with equal probability whatever the deleted keys. More generally, let us consider a sequence of operations $O_1O_2\ldots O_j$, the initial data structure being empty; any data type may be considered: linear list, priority queue, dictionary; assume O_j is the ith I of the sequence; let $x_1<x_2<\cdots<x_{i-1}$ be the keys inserted during the sequence $O_1O_2\ldots O_{j-1}$, and let w be the ith inserted key. Then, all the cases $w<x_1<x_2<\cdots<x_{i-1}$, $x_1<w<x_2<\cdots<x_{i-1}$, $\ldots$, $x_1<x_2<\cdots<x_{i-1}<w$ are equally likely, whatever the

deleted keys. Put into combinatorial language: after j operations in a linear list, out of which, say, i are I's, and, thus, $j-i$ are D's, the size of the data structure is $k=2i-j$; the keys of the data structure can be considered as a subset of k distinct objects of a set of size i, any of the $\binom{i}{k}$ possible subsets being equally likely.

We say that the number of possibilities of the ith I (in a sequence of operations) is equal to i (for Knuth's model) whatever the size of the data when this insertion occurs. On the contrary, in the Markovian model we say that the number of possibilities of an I_0 operation is $k+1$ iff k is the size of the data structure when this insertion occurs, whatever the history of the sequence and of the data structure. A similar proof can be given for Q^+ and Q^- (see [9] for more details). The possibility functions in Knuth's model are given in Table 1 (Section 3).

3. Basic definitions

(i) Following Flajolet et al. [5], we define a schema (or path) $Y(.)$, of length $2n$ as a word $\Omega := O_1 O_2 \ldots O_{2n} \in \{I, D, Q^+, Q^-\}^*$ such that for all j, $1 \leqslant j \leqslant 2n$

$$|O_1 O_2 \ldots O_j|_I \geqslant |O_1 O_2 \ldots O_j|_D. \qquad (0)$$

A schema is to be interpreted as a sequence of $2n$ requests (the keys operated on not being represented) where I, D, Q^+ and Q^- represent, respectively, an insertion, a deletion, a positive (successful) query (or search) and a negative (unsuccessful) query (or search). (0) means that the size of the structure is always $\geqslant 0$.

In the case of linear lists (LL) and priority queues (PQ), only insertions and deletions are performed. A *structure history* is a sequence of the form $h := O_1(r_1)O_2(r_2)\ldots O_{2n}(r_{2n})$, where $\Omega = O_1 O_2 \ldots O_{2n}$ is a schema, and

- r_j is the rank (or order) of the key operated upon at step j,
- $\alpha_j(\Omega) := |O_1 O_2 \ldots O_j|_I - |O_1 O_2 \ldots O_j|_D$ is the size (level) of the structure at step j.

We only consider schemas and histories with *initial and final level* 0 (the general case can be treated with similar techniques). The possibility function pos (defined for each request) is given in Table 1, where k denotes the size of the structure.

Let us return to the example given in Section 2. In terms of ranks, the history for the *Markovian model* can thus be derived as

I(0) I(1) D(0) I(0) D(0) I(0) D(0) D(0).

Table 1

Data type	pos(ith I)	pos(D, k)	pos(Q^+, k)	pos(ith Q^-)
Linear list (LL)	i	k	0	0
Priority queue (PQ)	i	1	0	0
Dictionary (D)	i	k	k	i

Table 2

	$\tau(h)$	
	SL	UL
LL	$\sum_{j\in(\mathrm{I}+\mathrm{D})} s_j(h)$	$\sum_{j\in\mathrm{D}} s_j(h)$
PQ	$\sum_{j\in\mathrm{I}} s_j(h)$	$\sum_{j\in\mathrm{D}} \alpha_j(h)$
D	$\sum_{j\in(\mathrm{I}+\mathrm{D}+\mathrm{Q}^+ +\mathrm{Q}^-)} s_j(h)$	$\sum_{j\in(\mathrm{D}+\mathrm{Q}^+)} s_j(h)+\sum_{j\in\mathrm{D}^-} \alpha_j(h)$

In *Knuth's model*, the history can be written as

$$\mathrm{I}(0)\ \mathrm{I}(1)\ \mathrm{D}(0)\ \mathrm{I}(0)\ \mathrm{D}(0)\ \mathrm{I}(3)\ \mathrm{D}(0)\ \mathrm{D}(0).$$

For any structure, let $\tilde{N}_{2n}$ be the (finite) *set* of histories of length $2n$ and let N_{2n} be the *number* of such histories (see [5]). For instance, N_{2n} is $n?$ for MPQ ($n? := 1 \cdot 3 \cdot 5 \cdots (2n-1)$).

(ii) To any history h, we will associate cost functions $C(h)$. Two cost functions are considered in this paper: the *storage cost function* $\sigma(h) := \sum_{j=1}^{2n} \alpha_j(h)$ and the *time cost function* $\tau(h)$. The latter function depends on the *implementation* of our lists' structures. We will use three implementations: the *sorted list* (SL), the *unsorted list* (UL) and the *binary tournament* (BT).

The time cost functions are summarized in Table 2, where s_j denotes the *position* of each key among all existing keys at time j. Let us explain the origin of these cost functions. For instance, for PQ in sorted list implementation (SL), we always delete the first (smallest) key: this costs just nothing. But, when we insert a key, we must first find its correct position: this costs s_j inspections.

As another example, let us consider D in unsorted list implementation (UL). Insertion obviously costs nothing: we just put the key in front of the list. Deletion and positive search require finding the position s_j of the key. An unsuccessful query needs to go through the entire list: this costs α_j inspections.

For the BT implementation, we only know the mean of an insertion: $H[\alpha_j(h)+1]-\frac{1}{2}$, and for a deletion: $2[H[\alpha_j(h)]-2+1/\alpha_j(h)]$ in a classical BT. (H(k) is the kth harmonic number.) Assume, for simplicity, that this remains true in Knuth's model (see the conjecture in Section 4.2.2).

Remark 1. For LL and PQ, we note that to each insertion at some level l, corresponds one deletion at level $l-1$.

With any cost function $C(h)$, we associate a random variable C^* defined as follows:

$$\Pr[C^*=\kappa]:=\frac{\operatorname{card}\{h:\ C(h)=\kappa,\ h\in\tilde{N}_{2n}\}}{N_{2n}}.$$

Expectation and variance of any event related to C^* are denoted by E^* and V^*. Following [19], we associate with each path $Y(.)$ a classical random walk (each step affected by the same probability) of length $2n$, from 0 to 0, with weight given by the possibility functions of Table 1.

Each trajectory will thus be affected by a total measure, which is the product of probability measures (related to large deviations) and a weight depending on the data type we consider. The *weighted* random walk corresponding to some path $Y(.)$ is denoted by $Y^*(.)$.

The next two sections are devoted to the characterization of the limiting distribution corresponding to each dynamic data structure (LL, PQ and D). These sections involve some difficult probabilistic tools. For the reader not familiar with these techniques, let us roughly explain the spirit of the method:

- First we have to prove a kind of central limit theorem for the process Y_n^*; this means that we have to find a centering term $\operatorname{Cent}(Y_n^*)$ and a normalizing term $\operatorname{Nor}(Y_n^*)$ such that $[Y_n-\operatorname{Cent}(Y_n^*)]/\operatorname{Nor}(Y_n^*)$ converges (weakly, as $n\to+\infty$) to a random process X.
- Then we must find the mean, the covariance and the distribution of X. Some results obtained previously by Louchard [19] are helpful here.

4. Limiting distributions

This section is organized in the following way: we only list the results for linear lists (LL) (Section 4.1), priority queues (PQ) (Section 4.2) and dictionaries (D) (Section 4.3) with a few direct derivations. The main steps of the proofs are given in Section 5. Some lemmas and theorems for D need very advanced techniques: the interested reader shall find them in the appendix.

4.1. The linear list in Knuth's model (KLL)

By [8] we see that KLL is combinatorially equivalent to the priority queue in Markovian model (MPQ), which we denote by KLL≡MPQ.

By [19, Theorem 5.3] we know that, for MPQ,

$$X_n(v):=\frac{Y^*([nv])-ny(v)}{\sqrt{n}}\Rightarrow X(v),\quad v\in[0,2],\tag{1}$$

where $y(v)=v(2-v)/2$, $X(.)$ is a Markovian Gaussian process with mean 0 and covariance $\gamma(x_1)\gamma(2-x_2)(x_1 \leqslant x_2)$, with $\gamma(v):=v^2/2$, and $\Rightarrow$ represents a weak convergence of random functions in the space of all right continuous functions having left limits and endowed with the Skorohod metric (see [1, Ch. III]). Also, it is proved in [19, Section 5.2] that

$$v_1:=\int_0^2 y(v)\,\mathrm{d}v=2/3, \quad v_2:=E\left[\left[\int_0^2 X(v)\,\mathrm{d}v\right]^2\right]=8/45,$$

$$v_3:=\int_0^2 y^2(v)\,\mathrm{d}v=4/15.$$

Storage and time cost functions are now analyzed by three theorems which exhibit Gaussian properties.

4.1.1. Storage cost σ^*_{KLL}

The storage cost σ^*_{KLL} is identical to σ^*_{MPQ}. By [19, Theorem 5.4] this gives the following theorem, where $\sim$ represents convergence in distribution (for $n\to\infty$) and $\mathcal{N}(M, V)$ is the normal (or Gaussian) random variable with mean M and variance V.

Theorem 1.

$$\frac{\sigma^*_{\mathrm{KLL}}-n^2v_1}{(n^3v_2)^{1/2}}=\frac{\sigma^*_{\mathrm{KLL}}-2n^2/3}{(8n^3/45)^{1/2}}\sim\mathcal{N}(0,1).$$

4.1.2. Time cost τ^*_{KLL}

The time cost depends on the implementation.

For the SL implementation, we have the following theorem.

Theorem 2.

$$\frac{\tau^*_{\mathrm{KLL,SL}}-n^2/3}{(n^3/15)^{1/2}}\sim\mathcal{N}(0,1).$$

For the UL implementation, we derive the following result.

Theorem 3.

$$\frac{\tau^*_{\mathrm{KLL,UL}}-n^2/6}{(n^3/45)^{1/2}}\sim\mathcal{N}(0,1).$$

4.2. The priority queue in Knuth's model (KPQ)

By [8] we know that KPQ is combinatorially equivalent to a Markov stack (MS): KPQ$\equiv$MS. The number of histories N_{2n} is C_{2n} for MS [5], where C_{2n}:=the nth

Catalan number $=\binom{2n}{n}/(n+1)$. The Markovian stack has been analyzed in [15–17]: it appears that

$$Y^*([2nv])/\sqrt{2n} \Rightarrow X^+(v), \tag{2}$$

where $X^+(v)$ is the standard Brownian excursion (BE) (see [2] for details on this process). Let us now proceed to the cost analysis, which yields the BE functionals.

4.2.1. Storage cost σ^*_{KPQ}

The storage cost σ^*_{KPQ} is identical to σ^*_{MS}. This has been analyzed in [17, Theorem 9]: this gives the following theorem.

Theorem 4.

$$\sigma^*_{\mathrm{KPQ}}/(2n)^{3/2} \sim \left[\int_0^1 X^+(v)\,\mathrm{d}v\right] \quad \textit{(Brownian excursion area).}$$

For instance,

$$E(\sigma^*_{\mathrm{KPQ}}) \sim \sqrt{\pi}\,n^{3/2}, \qquad E(\sigma^{*2}_{\mathrm{KPQ}}) \sim 10n^3/3. \tag{3}$$

Numerical distribution and moments of $[\int_0^1 X^+(v)\,\mathrm{d}v]$ *are given in* [16].

4.2.2. Time cost τ^*_{KPQ}

For the SL implementation we assume, for simplification, that for large structure size an inserted key is uniformly distributed among all existing keys. This assumption is presently under investigation. We then have the following limiting distribution.

Theorem 5.

$$\frac{\tau^*_{\mathrm{KPQ,SL}}}{(2n)^{3/2}} \sim \frac{1}{4}\left[\int_0^1 X^+(v)\,\mathrm{d}v\right].$$

For the UL implementation we derive, by Table 2 and Remark 1, that $\tau^*_{\mathrm{KPQ,UL}} \sim \frac{1}{2}\sigma^*_{\mathrm{KPQ}}$; hence, we have the following theorem.

Theorem 6.

$$\tau^*_{\mathrm{KPQ,UL}}/(2n)^{3/2} \sim \frac{1}{2}\left[\int_0^1 X^+(v)\,\mathrm{d}v\right].$$

For the BT implementation, we denote by ξ_i and η_i the random variables related to an insertion or a deletion in a BT of size $Y^*(i)$. We *conjecture* that, in Knuth's model and for large $Y^*(i)$,

$$E(\xi_i) \sim H(Y_i^*), \qquad E(\eta_i) \sim 2H(Y_i^*),$$

$$\mu_k(\xi_i) = \mathrm{O}[\log(Y^*(i))^{k/2}], \qquad \mu_k(\eta_i) = \mathrm{O}[\log(Y^*(i))^{k/2}], \tag{4}$$

where $\mu_k(Z)$ is the kth centered moment of Z. (This is proved in [18] for a classical binary search tree.)

With (4), we can deduce that

$$\frac{\tau^*_{\mathrm{KPQ,\,BT}} - 3n[\log(n-\gamma)]/2}{3n} \sim \left[\int_0^1 \log[X^+(v)]\,\mathrm{d}v\right]. \tag{5}$$

The moments of $[\int_0^1 X^+(v)\,\mathrm{d}v]$ have been given in [16]. Conjecture (4) and the moments of functional (5) will be analyzed in a forthcoming report.

4.3. The dictionary in Knuth's model (KD)

4.3.1. Limit Theorems

To obtain a formula like (1), we must first put a weight and a probability measure on the trajectory $y(.)$ (see [19] for details). The probability is deduced from [19, Eq. (38)]. The dominant term is given by

$$\exp\left[-n\int_0^2\left[\log(1-y'(v)^2)+y'(v)\log\left(\frac{1+y'(v)}{1-y'(v)}\right)\right]\mathrm{d}v\right]. \tag{6}$$

According to [9], the weight is given by

$$(S_1)!\exp[Z], \quad S_1 := [\#(\mathrm{I}) + \#(Q^-)], \quad Z := \sum_{i \in (\mathrm{D}+\mathrm{Q}^+)} \log\left(ny\left(\frac{i}{n}\right)\right). \tag{7}$$

This weight is more intricate than the classical ones used in [19]. The determination of $y(.)$ is first solved by the following theorem.

Theorem 7. $E^*[Y^*(\lceil nv \rceil)] \sim ny(v)$, *where* $y(.)$ *is given by the implicit equation*

$$(-1-\sqrt{\tfrac{2}{3}}\,y^{1/2})\sqrt{1-2\sqrt{\tfrac{2}{3}}y^{1/2}} = \begin{cases} v-1 & \textit{when } v<1, \\ 1-v & \textit{when } v>1. \end{cases}$$

The explicit solution is

$$y(v) := [\sqrt{\tfrac{3}{2}}(2\cos(\varphi/3)-1)/2]^{1/2}, \quad \varphi = \arccos[2v(2-v)-1]. \tag{8}$$

We must now analyze the limiting process $X(.)$. This is given by the following lemma.

Lemma 8. $X(.)$ *is a Markovian Gaussian process, with mean* 0 *and covariance*

$$C^*_{12}(x_1, x_2) = \gamma(x_1)\gamma(2-x_2), \quad x_1 \leqslant x_2,$$

where $\gamma(x) := [3+y'(x)][y'(x)-1]^3/8$.

We finally derive the complete limiting process, which is a superposition of two distinct Gaussian processes.

Theorem 9. *For KD,*

$$\frac{Y^*(\lceil nr \rceil - ny(r))}{\sqrt{n}} \Rightarrow X(r)+\mu(r),$$

where

- $y(.)$ *is given by* (8);
- $X(.)$ *is a Gaussian Markovian process with mean* 0 *and covariance* C^*_{12} *given by Lemma* 8;
- $\mu(x) := \int_0^2 \psi(x,r)\sqrt{s(y')}\,\mathrm{d}B(r)$, *where* $s(y') := [1-y'^2]/8$, *and* $\psi(x,r)$ *is given by Lemma* 16 (*see Section* 5) *and* $B(.)$ *represents a classical Brownian motion. This stochastic integral shows that* $\mu(x)$ *is a Gaussian (non-Markovian) process with covariance*

$$C\mu(x_1,x_2) := \int_0^2 \psi(x_1,r)\psi(x_2,r)s(y')\,\mathrm{d}r.$$

We can now analyze storage and time cost functions, which lead to Gaussian variables.

4.3.2. *Strorage cost* σ^*_{KD}

We obtain the following theorem.

Theorem 10.

$$\frac{\sigma^*_{\mathrm{KD}}-n^2\nu_1}{[n^3(\nu_2+\nu_4)]^{1/2}} \sim \mathcal{N}(0,1),$$

where $\nu_1 = 18/35$, $\nu_2 = 1656/13475$, *and* ν_4 *is given by* (9) *below.*

Proof. The result is deduced from Theorem 9 with

$$\nu_1 := \int_0^2 y(r)\,\mathrm{d}r,$$

$$\nu_2 := 2\int_0^2 \mathrm{d}u_1 \int_{u_1}^2 \mathrm{d}u_2\, C^*_{12}(u_1,u_2), \tag{9}$$

$$\nu_4 := 2\int_0^2 \mathrm{d}x_1 \int_{x_1}^2 \mathrm{d}x_2\, C\mu(x_1,x_2). \qquad \square$$

4.3.3. *Time cost* τ^*_{KD}

For the SL implementation, we have by Table 2

$$E^*(\tau^*_{\mathrm{KD,SL}}) \sim \tfrac{1}{2} n^2 v_1 = 9n^2/35.$$

Proceeding now as in Section 1.2, we readily obtain

$$V^*(\tau^*_{\mathrm{KD,SL}}) \sim \tfrac{1}{12} n^3 v_3 + \tfrac{1}{4} n^3 (v_2 + v_4),$$

with $v_3 := \int_0^2 y^2(v)\,\mathrm{d}v = 12/77$.

The Gaussian property of τ^*_{KD} is checked as in [19, Theorem 4.12]. We finally obtain the following Theorem.

Theorem 11.

$$\frac{\tau^*_{\mathrm{KD,SL}} - 9n^2/35}{[n^3(v_3/12 + (v_2+v_4)/4]^{1/2}} \sim \mathcal{N}(0,1).$$

For the UL implementation, we derive the following result.

Theorem 12.

$$\frac{\tau^*_{\mathrm{KD,UL}} - 17n^2/70}{V^*(\tau_{\mathrm{KD,UL}})^{1/2}} \sim \mathcal{N}(0,1).$$

Remark 2. It appears that the limiting distributions are mostly Gaussian; this is not an obvious (or trivial) fact since the classical results using central limit theorems are not directly applicable here. In addition, the limiting distributions for KPQ are not Gaussian: they depend on Brownian excursion functionals.

5. Main proofs' steps

5.1. *Proof of Theorems 2 and 3 for KLL*

For the SL implementation, we have from Table 2

$$\begin{aligned} E^*(\tau^*_{\mathrm{KLL,SL}}) &\sim \left[\sum_{Y^* \in \tilde{N}_{2n}} \sum_{i=1}^{2n-1} \left[\sum_{j_i=1}^{Y^*(i)} j_1 / Y^*(i) \right] \right] \Big/ n? \\ &\sim \left[\sum_{Y^* \in \tilde{N}_{2n}} \sum_{i=1}^{2n-1} \frac{Y^*(i)}{2} \right] \Big/ n? \\ &= \tfrac{1}{2} E^*(\sigma^*_{\mathrm{KLL}}) = \tfrac{1}{2} n^2 v_1 = n^2/3, \end{aligned}$$

$$V^*(\tau^*_{\mathrm{KLL,SL}}) \sim \left[\sum_{Y^* \in \tilde{N}_{2n}} \sum_{j_1=1}^{Y^*(1)} \cdots \sum_{j_{2n-1}=1}^{Y^*(2n-1)} \left[\sum_{i=1}^{2n-1} (j_i - E^*(Y^*(i))/2) \right]^2 \Big/ \prod_l Y^*(l) \right] \Big/ n?$$

$$\sim \left[\sum_{Y^* \in \tilde{N}_{2n}} \sum_{j_1=1}^{Y^*(1)} \cdots \sum_{j_{2n-1}}^{Y^*(2n-1)} \left[\sum_{1}^{2n-1} [j_i - \tfrac{1}{2} Y^*(i)] + \tfrac{1}{2} \sum_{1}^{2n-1} [Y^*(i) - E^*(Y^*(i))] \right]^2 \Big/ \prod_l Y^*(l) \right] \Big/ n?,$$

and by standard variance analysis (see the details in [19, Section 4.6.3])

$$V^*(\tau^*_{\mathrm{KLL,SL}}) \sim E^* \left[\tfrac{1}{12} \sum_{i=1}^{2n-1} Y^{*2}(i) \right] + \tfrac{1}{4} V^*(\sigma^*_{\mathrm{KLL}})$$

$$\sim \tfrac{1}{12} n^3 v_3 + \tfrac{1}{4} n^3 v_2 = n^3/15.$$

Repeating mutatis mutandis the proof of [19, Theorem 4.12], one finally obtains Theorem 2. □

For the UL implementation, we have from Table 2

$$E^*(\tau^*_{\mathrm{KLL,UL}}) \sim \left[\sum_{Y^* \in \tilde{N}_{2n}} \sum_{i \in \mathrm{D}} \left[\sum_{j_i=1}^{Y^*(i)} j_i / Y^*(i) \right] \right] \Big/ n?$$

$$= \tfrac{1}{4} E(\sigma^*_{\mathrm{KLL}}) \quad \text{(by Remark 1)} = \tfrac{1}{4} n^2 v_1 = n^2/6,$$

$$V^*(\tau^*_{\mathrm{KLL,UL}}) \sim \left[\sum_{Y^* \in \tilde{N}_{2n}} \sum \cdots \sum_{\substack{j_l=1 \\ l \in \mathrm{D}}}^{Y^*(l)} \cdots \sum \left[\sum_{i \in \mathrm{D}} j_i - \tfrac{1}{4} \sum_{i=1}^{2n-1} E^*(Y^*(i)) \right]^2 \Big/ \prod_{l \in \mathrm{D}} Y^*(l) \right] \Big/ n?$$

$$\sim E^* \left[\tfrac{1}{12} \sum_{i \in \mathrm{D}} Y^{*2}(i) \right]$$

$$+ E^* \left[\tfrac{1}{2} \left[\sum_{i \in \mathrm{D}} Y^*(i) - \tfrac{1}{2} \sum_{i=1}^{2n-1} E^*(Y^*(i)) \right] \right]^2$$

$$\sim \tfrac{1}{24} n^3 v_3 + \tfrac{1}{16} n^3 v_2 = n^3/45$$

The proof of [19, Theorem 5.6] can now be adapted; this leads to Theorem 3. □

5.2. *Proof of Theorem 5 for KPQ*

For the SL implementation, we have from Table 2 (with our simplifying assumption)

$$
\begin{aligned}
E^*(\tau^*_{\mathrm{KPQ.SL}}) &\sim \left[\sum_{Y^*\in\tilde{N}_{2n}} \sum_{i\in I} \frac{Y^*(i)}{2}\right] \Big/ C_{2n}\\
&\sim \tfrac{1}{2} E\left[2n \int_0^1 \sqrt{2n}\, \frac{X^+(v)}{2}\, \mathrm{d}v\right] && \text{[by (2) and Remark 1]}\\
&\sim \tfrac{1}{4}\sqrt{\pi}\, n^{3/2} && \text{[by (3)]},
\end{aligned}
$$

$$
\begin{aligned}
V^*(\tau^*_{\mathrm{KPQ.SL}}) \sim \Bigg[&\sum_{Y^*\in\tilde{N}_{2n}} \sum \cdots \sum_{\substack{j_l=1\\ l\in I}}^{Y^*(l)} \cdots \sum \Bigg[\Bigg(\sum_{i\in I} j_i - \tfrac{1}{4} \sum_{i=1}^{2n-1} Y^*(i) \Bigg)\\
&+ \tfrac{1}{2} \Bigg(\sum_{i\in I} Y^*(i) - \tfrac{1}{2} \sum_{i=1}^{2n-1} E^*(Y^*(i)) \Bigg) \Bigg]^2 \Big/ \prod_{l\in I} Y^*(l) \Bigg] \Big/ C_{2n}\\
\sim\ & \tfrac{1}{24}(2n)^2 E\left[\int_0^1 X^+(v)^2\, \mathrm{d}v\right] + (\tfrac{1}{4})^2 (2n)^3 V\left[\int_0^1 X^+(v)\, \mathrm{d}v\right],
\end{aligned}
\tag{10}
$$

where $V(Z)$ denotes the classical variance of Z.

From [14, p. 238],

$$
E\left[\int_0^1 X^+(v)^2\, \mathrm{d}v\right] = \int_0^\infty x^2 \cdot 4x\mathrm{e}^{-2x^2}\, \mathrm{d}x = \tfrac{1}{2}. \tag{11}
$$

We finally derive (the second term is *dominant* in (10))

$$
V^*(\tau^*_{\mathrm{KPQ.SL}}) \sim (\tfrac{1}{4})^2 (2n)^3 V\left[\int_0^1 X^+(v)\, \mathrm{d}v\right].
$$

More generally,

$$
\begin{aligned}
\mu_k[\tau^*_{\mathrm{KPQ.SL}}] &= E^*[\tau^*_{\mathrm{KPQ.SL}} - E^*(\tau^*_{\mathrm{KPQ.SL}})]^k\\
&\sim \sum_{r=0}^{k} \binom{k}{r} (\tfrac{1}{12} n^2)^{r/2} \mu_r[\mathcal{N}(0,1)] \left(\frac{(2n)^{3/2}}{4}\right)^{k-r}\\
&\quad \times \mu_{k-r}\left[\int_0^1 X^+(v)\, \mathrm{d}v\right].
\end{aligned}
\tag{12}
$$

Clearly, the dominant term of (12) is obtained with $r=0$. This gives

$$
\left(\frac{(2n)^{3/2}}{4}\right)^k \mu_k\left[\int_0^1 X^+(v)\, \mathrm{d}v\right];
$$

hence, Theorem 5. □

Our proof shows that, if the inserted key is *not* uniformly distributed among all existing keys, only the *numerical coefficient* 1/4 will be changed in Theorem 5.

For the *BT implementation*, we obtain from Table 2 [and noting that $H_k \sim \log(k)$]

$$\begin{aligned}
E^*(\tau^*_{\mathrm{KPQ,BT}}) &\sim \left[\sum_{Y^* \in \tilde{N}_{2n}} \left(\sum_{i\in I} \log(Y^*(i)\right) + \sum_{i \in D} 2\log(Y^*(i))\right] \Big/ C_{2n} \\
&\sim \frac{3}{2} E\left[2n \int_0^1 \log[\sqrt{2n}X^+(v)]\,\mathrm{d}v\right] \quad \text{[by Remark 1 and (2)]} \\
&= \frac{3n}{2}(\log 2 + \log n) + 3nE\left[\int_0^1 \log[X^+(v)]\,\mathrm{d}v\right] \\
&= \frac{3n}{2}(\log 2 + \log n) + 3n \int_0^{\infty} \log x \cdot 4x\mathrm{e}^{-2x^2}\,\mathrm{d}x \\
&= 3n\log(n/2) - 3\gamma n/2, \quad \text{[by (11)]}
\end{aligned}$$

$$\begin{aligned}
V^*(\tau^*_{\mathrm{KPQ,BT}}) \sim \Bigg[\sum_{Y^* \in \tilde{N}_{2n}} \hat{E} \Bigg[\Bigg[& \sum_{i \in I} (\xi_i - \log(Y^*(i))) + \sum_{i\in D} (\eta_i - 2\log(Y^*(i))) \Bigg] \\
& + \left[\sum_{i \in I} \log(Y^*(i)) + 2\sum_{i \in D} \log(Y^*(i))\right. \\
& \left. - \frac{3}{2}E^*(\log(Y^*(i)))\right]\Bigg]^2 \Bigg] \Big/ C_{2n}, \qquad (13)
\end{aligned}$$

where $\hat{E}$ is the expectation conditioned on Y^*. Then (13) becomes [conditioned on Conjecture (4)]

$$V^*(\tau^*_{\mathrm{KPQ,BT}}) \sim \left(\frac{3}{2}\cdot 2n\right)^2 V\left[\int_0^1 \log[X^+(v)]\,\mathrm{d}v\right].$$

More generally, we check that

$$\mu_k[\tau^*_{\mathrm{KPQ,BT}}]/(3n)^k \sim \mu_k\left[\int_0^1 \log[X^+(v)]\,\mathrm{d}v\right] \text{ and hence, (5).}$$

5.3. *Proofs of main lemmas and theorems for dictionaries*

The D case is more difficult to analyze; we have only given the main steps and a few proofs. More technical proofs are given in the appendix. To derive Theorem 7, we must first establish the probability of various steps *along* $ny(v)$. Let $p(\mathrm{I}, v) := P[\mathrm{step}[nv] \in \mathrm{I}]$ and similarly for D, Q^+, Q^-. We shall use the following lemma.

Lemma 13.

$$p(I,v)=\tfrac{1}{4}[1+2y'(v)+y'(v)^2]+\mathrm{O}\left(\frac{1}{n}\right),$$

$$p(D,v)=\tfrac{1}{4}[1-2y'(v)+y'(v)^2]+\mathrm{O}\left(\frac{1}{n}\right),$$

$$p(Q^-,v)=p[Q^+,v]=\tfrac{1}{4}[1-y'(v)^2]+\mathrm{O}\left(\frac{1}{n}\right).$$

The mean of this distribution is, of course, $y'(v)+\mathrm{O}(1/n)$.

The proof of Lemma 13 is given in the appendix.

We now need the total asymptotic measure along a path. We derive the following result.

Lemma 14. *The dominant term in the logarithm of the asymptotic total measure along* $ny(.)$ *is given by*

$$2n\log n-n+n\int_0^2\left[-\log(1-y'^2)-y'\log\left(\frac{1+y'}{1-y'}\right)+\frac{1}{2}(1-y')\log y\right]\mathrm{d}v$$

$$=2n\log n-n+n\int_0^2 f(y,y')\,\mathrm{d}v,\ \text{say.} \tag{14}$$

The proof is also given in the appendix. We can now turn to the determination of $y(.)$.

Proof of Theorem 7. Maximizing (14) is a variational problem, which can be solved as in [19, Section 4.4]. This gives the equation

$$1-y'^2=C_1\sqrt{y}, \tag{15}$$

the implicit solution of which is

$$(-8/3C_1^2-4y^{1/2}/3C_1)(1-C_1y^{1/2})^{1/2}=\begin{cases} v+C_2 & \text{when } y'>0\\ -C_2-v & \text{when } y'<0.\end{cases} \tag{16}$$

The constraints $y(0)=y(2)=0$ lead to $C_1=2(2/3)^{1/2}$, $C_2=-1$. The explicit form (8) is given by the suitable solution of the cubic equation corresponding to (16). □

Proof of Lemma 8. To find the distribution of $X(x)$ [see (1)], we must first include the contribution from (14). Letting $\theta:=X_n(x)/\sqrt{n}$, this can be deduced from [19, Lemma 4.7]; it amounts to

$$n\frac{\theta^2}{2}\int_0^x[f_{yy}z_1^2+2f_{yy'}z_1z_1'+f_{y'y'}(z_1'^2)]\,\mathrm{d}v\ +n\int_x^2[\text{symmetric term}]\,\mathrm{d}v,$$

where $z(v)$ is given by (16) with

- $u := y(x) + \theta$,
- $z(0) = 0,\ z(x) = u,$ (17)
- $z_1 := \left(\frac{\partial z}{\partial u}\right)_{u=y(x)},\ z_1' := \frac{\partial z_1}{\partial v}.$

The detailed techniques to obtain z, z_1, z_1' can be found in [19, Sections 4.5.1 and 4.5.2].

Actually, we will not use the complicated expression (8). All integrations will be performed with respect to a new variable $h = (1 - 2\sqrt{2/3}\,y^{1/2})^{1/2}$, which, by (15) is nothing but $|y'|$. The first integral leads to

$$-\int_0^x \left[\frac{z_1^2(1-y')}{2y^2} + \frac{z_1 z_1'}{y} + \frac{2z_1'^2}{1-y'^2}\right] \mathrm{d}v.$$

Performing this integration and adding the term from $\int_x^2$ leads to $(y' \equiv y'(x))$

$$\frac{-n\theta^2}{2} 64/[(-1+y')^3(y'-3)(y'+1)^3(y'+3)],$$

which shows that $X(x)$ is a *Gaussian variable* with mean 0 and variance

$$V^*(x) = \gamma(x)\gamma(2-x), \quad \gamma(x) = [3 + y'(x)][y'(x) - 1]^3/8.$$

The covariance can be obtained by similar computation (we omit the details). □

Proof of Theorem 9. Let the (normalized) trajectory be

$$z(v) := y(v) + \chi(v)/\sqrt{n} \quad (\chi(0) = \chi(2) = 0).$$

We must now determine the contribution of χ, χ' arising from (7). This is given by the following lemma, where we obtain stochastic integrals on Brownian motion.

Lemma 15. *The linear contribution to* χ, χ' *from* $\log[(S_1)!] \cdot Z$ *is asymptotically given by*

$$\chi(v)\frac{\sqrt{s(y')}}{y(v)}\mathrm{d}B(v) + \chi'(v)\frac{\log(y(v))}{2\sqrt{s(y')}}\frac{\partial s}{\partial y'}\mathrm{d}B(v)$$
$$= \chi(v)\,\mathrm{d}h_1(v) + \chi'(v)\,\mathrm{d}h_2(v),\ \textit{say}, \tag{18}$$

where $B(.)$ *is a classical Brownian motion and* $s(y') := [1 - y'^2]/8$.

The proof is given in the appendix. Lemma 15 (which gives the contribution of (7) to our process density) tells us that we must transform our expression for $X(.)$ into

$X(v):=\chi(v)-\mu(v)$ and we must determine $\mu(.)$. This is given by the following lemma, the proof of which is also given in the appendix.

Lemma 16. *$\mu(v)=\mu_1(v)+\mu_2(v)$, where*

$$\mu_1(v)=\gamma(v)\int_v^2 \gamma(2-u)\,\mathrm{d}h_1(u)+\gamma(2-v)\int_0^v \gamma(u)\,\mathrm{d}h_1(u), \tag{19}$$

$$\mu_2(v)=-\gamma(v)\int_v^2 \gamma'(2-u)\,\mathrm{d}h_2(u)+\gamma(2-v)\int_0^v \gamma'(u)\,\mathrm{d}h_2(u), \tag{20}$$

and $h_1(.)$, $h_2(.)$ are given by Lemma 15. As expected, $\mu(0)=\mu(2)=0$. For further use, let us write $\mu(v)$ as $\int_0^2 \psi(v,u)\sqrt{s(y')}\,\mathrm{d}B(u)$.

Collecting now the results from (8) and Lemmas 8, 15 and 16, we readily obtain Theorem 9.

6. Conclusion

We have analyzed asymptotic distributions of linear lists, priority queues, and dictionaries, histories and cost functions in Knuth's model. It appears that the limiting cost distributions are either Gaussian random variables or Brownian excursion functionals. The limiting processes are Gaussian Markovian, Brownian excursion and Gaussian (non-Markovian) stochastic integrals. As further work, we intend to consider the symbol table in both Markovian and Knuth's model.

Appendix for KD

This appendix contains the detailed technical proofs of some lemmas and theorems related to KD.

Proof of Lemma 13. By [19, Lemma 4.1] (adapted to the case of dictionaries), we obtain by the theory of large deviations

$$\begin{aligned} P[\,Y(nt)\in n\,\mathrm{d}x] &\sim \left[\left(1-\frac{x^2}{t^2}\right)^{-(nt+1/2)}\left(\frac{1+x/t}{1-x/t}\right)^{-nx}\right]\frac{\sqrt{n}\,\mathrm{d}x}{\sqrt{\pi}\sqrt{t}} \\ &=\varphi(x,t)\,\mathrm{d}x, \text{ say,} \end{aligned}$$

where $Y(.)$ is the classical random walk related to dictionaries. Hence,

$$P[\text{step}[1]\in I\mid Y(nt)\in n\,\mathrm{d}x]\sim\tfrac{1}{4}\left[1+\left(\varphi\left(x-\frac{1}{n},t-\frac{1}{n}\right)-\varphi(x,t)\right)\Big/\varphi(x,t)\right].$$

After some tedious but simple manipulations, this gives

$$\tfrac{1}{4}[1+2x/t+(x/t)^2]+\mathrm{O}\left(\frac{1}{n}\right),$$

and, similarly, for D, Q^+, Q^-. One could also prove these formulas by starting directly from the dictionary large deviation generating function (see [19, (9)]). Proceeding now as in [19, Section 4.4], the lemma is easily proved.

Proof of Lemma 14. Let

$$p(y')=p(\mathrm{I},r)+p(\mathrm{Q}^-,r)=\tfrac{1}{2}[1+y'(r)]+\mathrm{O}\left(\frac{1}{n}\right),$$

$$q(y')=1-p(y')=p(D,r)+p(\mathrm{Q}^+,r)=\tfrac{1}{2}[1-y'(r)]+\mathrm{O}\left(\frac{1}{n}\right). \tag{21}$$

We see that the dominant term of S_1 along $ny(r)$ is given by

$$n\int_0^2 p(y')\,\mathrm{d}r=n+\mathrm{O}(1).$$

and

$$\begin{aligned}Z&\sim n\int_0^2 q(y')\log[ny(r)]\,\mathrm{d}r\\&=n\left[\log n+\mathrm{O}\left[\frac{\log n}{n}\right]+\int_0^2 \tfrac{1}{2}(1-y'(r))\log(y(r))\,\mathrm{d}r+\mathrm{O}\left(\frac{1}{n}\right)\right].\end{aligned} \tag{23}$$

By Stirling's formula, the dominant *variable* [in $y(.)$] part of (7) is immediately deduced from (22) and (23); this gives only

$$\exp\left[n\tfrac{1}{2}\int_0^2(1-y')\log y\,\mathrm{d}r\right] \tag{24}$$

(dropping the (r) to simplify formulas). Collecting the results from (6) and (24) we obtain Lemma 14.

To prove Lemma 15, we must take into account the stochastic part of (7). This is a new problem that we did not encounter in the more classical structures of [19]: only dominant terms like (14) were necessary there. To solve this problem, we first establish the *joined distribution of* S_1 *and* Z in the following lemma.

Lemma 17. *S_1 and Z along $nz(.)$ are asymptotically Gaussian, with*

$$E(S_1)=n\alpha_1,\qquad E(Z)=n\alpha_3,$$

$$\operatorname{cov}(S_1,Z)=\begin{pmatrix} n\alpha_2 & -n\alpha_5\\ -n\alpha_5 & n\alpha_4\end{pmatrix},$$

where $z' \equiv z'(v)$, and

$$\alpha_1 := \int_0^2 p(z')\,\mathrm{d}v = 1 + \mathrm{O}\left(\frac{1}{n}\right),$$

$$\alpha_2 := \int_0^2 r(z')\,\mathrm{d}v, \quad r(z') := s(z') + \mathrm{O}\left(\frac{1}{n}\right), \quad s(z') := [1 - z'^2]/8,$$

$$\alpha_3 := \int_0^2 q(z')\log[nz]\,\mathrm{d}v,$$

$$\alpha_4 := \int_0^2 r(z')\log^2[nz]\,\mathrm{d}v,$$

$$\alpha_5 := \int_0^2 r(z')\log[nz]\,\mathrm{d}v.$$

Proof. Let us divide the $2n$ steps into $2n/m$ groups of m steps, where m is large and

$$m = \mathrm{o}(n). \tag{25}$$

In each group, we must study the asymptotic distribution of S_1 and $S_2 := m - S_1$. This can be done as follows. To fix a group, let $t_1, t_2 \in [0, 2]$, $t_2 = t_1 + \Delta$, $z_1 = z(t_1)$, $z_2 = z(t_2)$, and $m = n\Delta$.

If we *do not constrain* the random walk $z(.)$ associated with the dictionary to follow, we can use the probabilities given by Lemma 13 to obtain, in $[nt_1, nt_2]$,

$$\#(\mathrm{I}) \sim mp(\mathrm{I}) + \eta(\mathrm{I}), \quad \eta(I) = \mathcal{N}(0, mp(I)q(\mathrm{I})), \; q(\mathrm{I}) := 1 - p(\mathrm{I}),$$

and, similarly, for $\mathrm{D}, \mathrm{Q}^-, \mathrm{Q}^+$.

We have dropped the v dependence as $z'(.)$ is asymptotically constant in $[t_1, t_2]$ by (25).

The covariance of $(\#(\mathrm{I}), \#(\mathrm{Q}^+))$ is easily seen to be $-mp(\mathrm{I})p(\mathrm{Q}^+)$ and, similarly, for $(\mathrm{I}, \mathrm{Q}^+, \mathrm{Q}^-, \mathrm{D})$. The covariance matrix is obviously of rank 3, which reflects the fact that

$$\#(\mathrm{I}) + \#(\mathrm{D}) + \#(\mathrm{Q}^+) + \#(\mathrm{Q}^-) = m, \quad \text{i.e.} \quad \eta(\mathrm{I}) + \eta(\mathrm{Q}^-) + \eta(\mathrm{Q}^+) + \eta\,(\mathrm{D}) = 0. \tag{26}$$

This allows us to represent all random variables as functions of $[\eta(\mathrm{I}), \eta(\mathrm{Q}^-), \eta(\mathrm{Q}^+)]$ with covariance matrix C given by

$$C = m \begin{array}{c} \left(\begin{array}{ccc} p(\mathrm{I})\cdot q(\mathrm{I}) & -p(\mathrm{I})\cdot p(\mathrm{Q}^-) & -p(\mathrm{I})\cdot p(\mathrm{Q}^+) \\ -p(\mathrm{Q}^-)\cdot p(\mathrm{I}) & p(\mathrm{Q}^-)\cdot q(\mathrm{Q}^-) & -p(\mathrm{Q}^-)\cdot p(\mathrm{Q}^+) \\ -p(\mathrm{Q}^+)\cdot p(\mathrm{I}) & -p(\mathrm{Q}^+)\cdot p(\mathrm{Q}^-) & p(\mathrm{Q}^+)\cdot q(\mathrm{Q}^+) \end{array}\right) \end{array} \begin{array}{l} \mathrm{I} \\ \mathrm{Q}^- . \\ \mathrm{Q}^+ \end{array}$$

The density of $[\eta(\mathrm{I}), \eta(\mathrm{Q}^-), \eta(\mathrm{Q}^+)]$ is characterized by the matrix C^{-1}. Now, if we *constrain* the random walk $Y(.)$ on $[nt_1, nt_2]$ to be such that $Y(nt_1)=nz_1$, $Y(nt_2)=nz_2$, this amounts to imposing the relation

$$\eta(\mathrm{I})\Delta(\mathrm{I})+\eta(\mathrm{Q}^-)\Delta(\mathrm{Q}^-)+\eta(\mathrm{Q}^+)\Delta(\mathrm{Q}^+)+\eta(\mathrm{D})\Delta(\mathrm{D})=0, \tag{27}$$

where $\Delta(.)$ are the four increments, centered on their mean z', i.e. $\Delta(\mathrm{I})=1-z'$, $\Delta(\mathrm{Q}^-)=\Delta(\mathrm{Q}^+)=-z'$, $\Delta(\mathrm{D})=-1-z'$.

Inserting (26) into (27) gives

$$\eta(\mathrm{Q}^+)=-2\eta(\mathrm{I})-\eta(\mathrm{Q}^-).$$

This last equation is now inserted into C^{-1} giving the matrix A^{-1} corresponding to the density of the constrained couple $(\eta(\mathrm{I}), \eta(\mathrm{Q}^-))$. Actually, J. Leroy has shown to us that a suitable simple matrix transformation, applied to C, easily leads to A. We obtain

$$A=m\begin{array}{c} \\ \left(\begin{array}{cc} 1/16-y'^2/8+y'^4/16 & -1/16+y'^2/8-y'^4/16 \\ -1/16+y'^2/8-y'^4/16 & (y'^2-3)(y'^2-1)/16 \end{array}\right)\end{array}\begin{array}{l}\mathrm{I}\\ \mathrm{Q}^-\end{array}.$$

We immediately deduce, from the distribution of $\eta(\mathrm{I})+\eta(\mathrm{Q}^-)$ on a group, that

$$S_1\sim mp(z')+\mathcal{N}(0, mr(z')), \quad r(z')=(1-z'^2)/8+\mathrm{O}(1/m),$$

We also have

$$S_2=[\#(\mathrm{Q}^+)+\#(\mathrm{D})]=m-S_1.$$

$\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are now immediately deduced as all groups are independent by constraint (27). To derive α_5, note that, on $[nt_1, nt_2]$, $E[(S_1-mp(z'))(S_2-m(1-p(z'))]=-mr(z')$; hence, α_5 follows. □

Proof of Lemma 15. Let us return to the techniques we used in Lemma 17: the $2n$ steps are divided into $n_g := 2n/m$ groups of m steps. In group j $(j=1,\ldots,n_g)$, let

$$S_{1,j} := (\#_j(\mathrm{I})+\#_j(\mathrm{Q}^-))=mp(z_j')+[\eta_j(\mathrm{I})+\eta_j(\mathrm{Q}^-)]=mp(z_j')+\omega_j, \text{ say,}$$

where, by the proof of Lemma 17 $\omega_j=\mathcal{N}(0, mr(z_j'))$ and z_j' is the value of $z'(v)$ on group j [asymptotically constant on that group by (25)].

Now, from Lemma 17 again, we obtain

$$S_1\sim n\alpha_1+\sum_{j=1}^{n_g}\omega_j, \qquad Z\sim n\alpha_3+\sum_{j=1}^{n_g}[-\log(nz_j)\omega_j].$$

It is now a classical expansion exercise to obtain the χ, χ' term from $\log[(S_1)!]\cdot Z$. We first derive

$$\log[(S_1)!]\cdot Z \sim [-n\alpha_1 + n\alpha_1 \log(n\alpha_1) + n\alpha_3]$$
$$+\left[\left(\sum_{j=1}^{n_g} \omega_j\right)\log(n\alpha_1) + \sum_{j=1}^{n_g} [-\log(nz_j)\omega_j]\right].$$

The term in the first square brackets on the right-hand side is, of course, the $O(n)$ contribution to (14). The term in the second square brackets, by Lemma 17, reduces to

$$\sum_{j=1}^{n_g} [-\log(z_j)\omega_j]. \tag{28}$$

Note that $-\omega_j$ [which can also be written as $\eta_j(D)+\eta_j(Q^+)$] can be represented by $\sqrt{mr(z'_j)}\,\xi_j$, where $\xi_j := \mathcal{N}(0,1)$ and all ξ_j are independent. Expanding (28) we derive for the linear χ, χ' term

$$\frac{\sqrt{m}}{\sqrt{n}} \sum_{j=1}^{n_g} \left[\frac{\chi_j}{y_j}\sqrt{s_j} + \log(y_j)\frac{\partial s}{\partial y'_j}\frac{\chi'_j}{2\sqrt{s_j}}\right]\xi_j,$$

where $s_j := s(y'_j)$ and, similarly, for χ_j, χ'_j, y_j.

Setting now $\Delta v := m/n\,(v\in[0,2])$, we formally obtain

$$\sum_{j=1}^{n_g} \left[\frac{\chi_j}{y_j}\sqrt{s_j} + \log(y_j)\frac{\partial s}{\partial y'_j}\frac{\chi'_j}{2\sqrt{s_j}}\right]\xi_j\sqrt{\Delta v_j},$$

where we recognize a classical Gaussian white noise: it is well known that $\xi_j\sqrt{\mathrm{d}v_j}$ can be written as $\mathrm{d}B(v_j)$; hence, (18) follows. □

Proof of Lemma 16. Let a family of Gaussian variables $x=(x_1 \dots x_k)$ characterized by a density $\exp[-\frac{1}{2}Q(x)]/(2\pi)^{k/2}|Q|^{-1}$, where the quadratic form $Q(x)$ is constructed from the matrix Q. If we set $x=\rho-\lambda$, where λ is the mean of the Gaussian family ρ, the linear term (in ρ) of the logarithm of this density is given by $\rho Q\lambda^{\mathrm{T}}$. If we know that this linear term is given by ρg^{T}, we derive $g^{\mathrm{T}} = Q\lambda^{\mathrm{T}}$; hence, $\lambda^{\mathrm{T}} = Q^{-1}g^{\mathrm{T}}$. But it is well known that $Q^{-1}\equiv C$, where C is the covariance of x, i.e. $C=E[\,x^{\mathrm{T}}x]$; hence,

$$\lambda^{\mathrm{T}} = C_g^{\mathrm{T}} \tag{29}$$

In our case, we deal with continuous time processes and we must use, for the $h_1(.)$ contribution, the correspondence relations:

$$x_i \leftrightarrow X(v),\quad \rho_i \leftrightarrow \chi(v),\quad \lambda_i \leftrightarrow \mu_1(v),\ C_{ij} \leftrightarrow \gamma(v)\gamma(2-u),\qquad v\leqslant u,$$
$$g_i \leftrightarrow \mathrm{d}h_1(v).$$

Equation (19) is now immediate from (29).

For the $h_2(.)$ contribution, we could be tempted to formally transform $\int_0^2\chi'(v)\,\mathrm{d}h_2(v)$ into $-\int_0^2\chi(x)(h''_2(v)\,\mathrm{d}v$ ($\chi(0)=\chi(2)=0$) but we cannot, of course, differentiate the Brownian contribution to $\mathrm{d}h_2$. So, we turn to another technique: assume that we have a contribution $\int_0^2\chi(u)w(u)\,\mathrm{d}u$ to the density for some $w(.)$. This is equivalent to

$$-\int_0^2\chi'(u)W(u)\,\mathrm{d}u, \tag{30}$$

where $W(u):=\int_0^u w(s)\,\mathrm{d}s$. Expression (30) can also be written as

$$-\int_0^2\chi'(v)\,\mathrm{d}\mathscr{W}(u),\ \text{where}\ \mathscr{W}(u):=\int_0^u W(s)\,\mathrm{d}s. \tag{31}$$

But (19) gives here

$$\mu(v)=\gamma(v)\int_v^2\gamma(2-u)w(u)\,\mathrm{d}u+\gamma(2-v)\int_0^v\gamma(u)w(u)\,\mathrm{d}u,$$

which can be transformed into

$$\mu(v)=-\gamma(v)\int_v^2\gamma'(2-u)\,\mathrm{d}\mathscr{W}(u)+\gamma(2-v)\int_0^v\gamma'(u)\,\mathrm{d}\mathscr{W}(u).$$

Comparing (31) with Lemma 15, we see that here $-\mathrm{d}\mathscr{W}(u)\equiv\mathrm{d}h_2(u)$; hence, (20) follows. □

Remark 3. Another way of proving (20) is to return to Lemma 8: we see that $X(.)$ can be written as $X(v)=\gamma(2-v)B(\gamma(v)/\gamma(2-v))$ for some Brownian motion $B(.)$. We easily deduce that

$$E[\mathrm{d}X(v_1)\,\mathrm{d}X(v_2)]=-\gamma'(v_1)\gamma'(2-v_2)\,\mathrm{d}v_1\,\mathrm{d}v_2,\quad v_1<v_2, \tag{32}$$

$$E[\mathrm{d}X^2(v_1)]=[\gamma(2-v_1)\gamma'(v_1)+\gamma'(2-v_1)\gamma(v_1)]\,\mathrm{d}v_1. \tag{33}$$

But dealing now with the process $\mathrm{d}X(v)$, we have $\mathrm{d}X(v)=\mathrm{d}\chi(u)-\mathrm{d}\mu(v)$. Using (29), (30), (32) and (33), we deduce that

$$\begin{aligned}\mathrm{d}\mu(v)={}&[-\gamma'(2-v)\gamma(v)-\gamma'(v)\gamma(2-v)]\,W(v)\,\mathrm{d}v\\&+\left[\gamma'(2-v)\int_0^v\gamma'(u)W(u)\,\mathrm{d}u+\gamma'(v)\int_v^2\gamma'(2-u)W(u)\,\mathrm{d}u\right]\mathrm{d}v.\end{aligned} \tag{34}$$

After a few elementary manipulations, and setting again $\mathrm{d}\mathscr{W}(u)=W(u)\,\mathrm{d}u$, we derive (20) from (34).

Proof of Theorem 12. By Table 2 and Lemma 13, we derive that

$$E^*(\tau^*_{\mathrm{KD,UL}}) \sim n^2 \left[\tfrac{1}{2} \int_0^2 p(\mathrm{D}, v) y(v)\, \mathrm{d}v + \tfrac{3}{2} \int_0^2 p(\mathrm{Q}^+, v) y(v)\, \mathrm{d}v \right] = 17n^2/70.$$

To analyze the variance, it is more convenient to return to the Lemma 15 representation. This gives us from Table 2

$$\begin{aligned} V^*(\tau^*_{\mathrm{KD,UL}}) \sim E \Bigg\{ \sum_{j=1}^{n_{\mathrm{g}}} \sum_{i=1}^{mq(z'_j)-\omega_j} \xi_{i,j} + (nz_j)[mp(\mathrm{Q}^-, z'_j) + \eta_j(\mathrm{Q}^-)] \Bigg] \\ - \sum_{j=1}^{n_{\mathrm{g}}} (ny_j) m \left[\frac{q(y'_j)}{2} + p(\mathrm{Q}^-, y'_j) \right] \Bigg\}^2, \end{aligned} \tag{35}$$

where

- $-\omega_j := [\eta_j(\mathrm{D}) + \eta_j(\mathrm{Q}^+)]$,
- $\xi_{i,j}$ are independent, uniformly $[0\,.\,.nz_j]$ distributed random variables
- $p(\mathrm{Q}^-, z'_j)$ is given by Lemma 13 and $q(.)$ by (21).

Using the techniques of Lemma 17, it is easily seen that

$$\eta(\mathrm{Q}^-) = -2\eta(\mathrm{D}) - \eta(\mathrm{Q}^+), \qquad \eta(\mathrm{I}) = \eta(\mathrm{D}).$$

Let $\varphi(z') := \frac{1}{2} q(z') + p(\mathrm{Q}^-, z') = \frac{1}{2} - \frac{1}{4} z' - \frac{1}{4} z'^2$ by (21) and Lemma 13. Developing (35) by classical variance analysis, we derive

$$\begin{aligned} V^*(\tau^*_{\mathrm{KD,UL}}) = E \Bigg\{ \sum_{j=1}^{n_{\mathrm{g}}} \Bigg[\sum_{i=1}^{mq(z'_j)-\omega_j} \xi_{i,j} - \tfrac{1}{2}(nz_j)[mq(z'_j) - \omega_j] \Bigg] \\ + \sum_{j=1}^{n_{\mathrm{g}}} (nz_j)[-\tfrac{3}{2}\eta_j(\mathrm{D}) - \tfrac{1}{2}\eta_j(\mathrm{Q}^+)] \\ + \sum_{j=1}^{n_{\mathrm{g}}} (nz_j) m [\varphi(z'_j) - \varphi(y'_j)] \\ + \sum_{j=1}^{n_{\mathrm{g}}} m\varphi(y'_j) n(z_j - y_j) \Bigg\}^2. \end{aligned} \tag{36}$$

In the last two brackets, we have, by Theorem 9,

$$n(z_j - y_j) \sim \sqrt{n}[X(x_j) + \mu(x_j)], \quad z'_j - y'_j \sim \frac{1}{\sqrt{n}}[X'(x_j) + \mu'(x_j)],$$

where group j is characterized by the interval $[x_j, x_{j+1}]$.

Note that, by (28), $\mu(x)$ is asymptotically given by

$$\mu(x_j) \sim \frac{1}{\sqrt{n}} \sum_{k=1}^{n_{\mathrm{g}}} [\psi(x_j, v_k)(\eta_k(\mathrm{D}) + \eta_k(\mathrm{Q}^+))].$$

Keeping only the dominant terms in (36), we finally derive (omitting the details)

$$V^*(\tau^*_{\mathrm{KD,UL}}) \sim n^3 \left[\tfrac{1}{12} \int_0^2 y^2 q(y')\,\mathrm{d}v \right.$$

$$+ V \left[\int_0^2 X(v)\varphi(y')\,\mathrm{d}v + \int_0^2 X'(v) \frac{\partial \varphi(y')}{\partial y'} y\,\mathrm{d}v \right]$$

$$\left. + \int_0^2 \tilde{r}(y, y')\,\mathrm{d}v \right], \tag{37}$$

where

$\tilde{r}(y, y') = t\tilde{A}t^{\mathrm{T}}$ ($\tilde{A}$ is given in the proof of Lemma 11 by $\tilde{A} = A/m$),

$t(v) := (-\frac{1}{2}y(v) - g(v), \frac{1}{2}y(v) - g(v))$,

$$g(v) := \left[\int_0^2 \varphi(y'(x))\psi(x, v)\,\mathrm{d}x + \int_0^2 \frac{\partial \varphi(y')}{\partial y'} \psi'(x, v)v(x)\,\mathrm{d}x \right],$$

where $\psi'(x, v) := \partial \psi(x, v)/\partial x$. The detailed computation of v_4 [see (9)] and $V^*(\tau^*_{\mathrm{KD,UL}})$ is under investigation. Note that (32) and (33) must be used for the middle term of (37).

By central limit theorem techniques (we omit the details, see [19, Section 4.6.3]), we can now derive Theorem 12. □

Acknowledgment

The authors are grateful to J. Françon for asking some questions which led to this work and to Ph. Flajolet, P. Lescanne, J.L. Rémy and J.M. Steyaert for several discussions on this topic. All complicated computations in this paper were done with the help of MAPLE. The authors express their gratitude to the Brussels University CSO unit for its help on several occasions.

The pertinent comment of our referees led to substantial improvement of the presentation.

References

[1] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).

[2] K.L. Chung, Excursions in Brownian motion, *Ark. Mat.* **14** (1976) 155–177.

[3] Ph. Flajolet, Analyse d'algorithmes de manipulation d'arbres et de fichiers, in: *Cahiers du BURO* **34–35** (1981).

[4] Ph. Flajolet, J. Françon and J. Vuillemin, Sequence of operations analysis for dynamic data structures, *J. Algorithms* **1** (1980) 111–141.

[5] Ph. Flajolet, C. Puech and J. Vuillemin, The analysis of simple lists structures, *Inform. Sci.* **38** (1986) 121-146.

[6] J. Françon, Histoire de fichiers, *RAIRO Inform. Theor.* **12** (1978) 49–62.

[7] J. Françon, Combinatoire des structures de données. Thèse de doct. d'Etat. Université de Strasbourg, 1979.

[8] J. Françon, B. Randrianarimanana and R. Schott, Analysis of dynamic data structures in D.E. Knuth's model, *Theoret. Comput. Sci.* **72** (1990) 147–167.

[9] J. Françon, B. Randrianarimanana, R. Schott, Analysis of dynamic algorithms in D.E. Knuth's model, in *Proc. CAAP '88*, Lecture Notes in Computer Science, Vol. 299 (Springer, Berlin, 1988) 72–88.

[10] A. Jonassen and D.E. Knuth, A trivial algorithm whose analysis isn't, *J. Comput. System Sci.* **16** (1978) 301–332.

[11] G.D. Knott, Deletion in binary storage trees, Report Stan-CS 75–491, 1975.

[12] D.E. Knuth, *The Art of Computer Programming, Vol. 3: Sorting and Searching* (Addison-Wesley, Reading, MA, 2nd ed., 1975).

[13] D.E. Knuth, Deletions that preserve randomness, *IEEE Trans. Software Engrg.* **3**(5) (1977) 351–359.

[14] P. Levy, *Processus Stochastiques et Mouvement Brownien* (Gauthier-Villars, 1948).

[15] G. Louchard, Kac's formula, Levy's local time and Brownian excursion. *J. Appl. Probab.* **21** (1984) 479–499.

[16] G. Louchard, The Brownian excursion area: a numerical analysis, *Comput. Math. Appl.* **10**(6) (1986) 413–417.

[17] G. Louchard, Brownian motion and algorithms complexity, *BIT* **26** (1986) 17–34.

[18] G. Louchard, Exact and asymptotic distributions in digital and binary search trees, *Theor. Inf. Appl.* **21**(4) (1987) 479–496.

[19] G. Louchard, Random walks, Gaussian processes and list structures, *Theoret. Comput. Sci.* **53** (1987) 99–124.

[20] G. Louchard, B. Randrianarimanana, R. Schott, Dynamic algorithms in D.E. Knuth's model: a probabilistic analysis, in: *Proc. ICALP '89*, Lecture Notes in Computer Science, Vol. 372 (Springer, Berlin, 1989) 321–333.

[21] B. Randrianarimanana, Analyse des structures de données dynamiques dans le modèle de D.E. Knuth, Thèse de 3ème Cycle, Université de Nancy 1, 1986.