RAIRO. INFORMATIQUE THÉORIQUE

G. LOUCHARD

The brownian motion : a neglected tool for the complexity analysis of sorted tables manipulation

RAIRO. Informatique théorique, tome 17, n° 4 (1983), p. 365-385

<http://www.numdam.org/item?id=ITA_1983__17_4_365_0>

R.A.I.R.O. Informatique théorique/Theoretical Informatics
(vol. 17, n° 4, 1983, p. 365 à 385)

THE BROWNIAN MOTION: A NEGLECTED TOOL FOR THE COMPLEXITY ANALYSIS OF SORTED TABLES MANIPULATION (*)

by G. LOUCHARD (1)

Communicated by J.-F. PERROT

Abstract. — *The Brownian Motion is shown to be a useful tool in analysing sorted tables:*
— firstly to easily get asymptotic results on the complexity of manipulation algorithms;
— secondly to shed more light on the probabilistic behaviour of these algorithms.

Résumé. — *Dans cet article, le Mouvement Brownien est proposé comme outil efficace dans l'analyse des tables ordonnées :*
— d'abord pour obtenir aisément des résultats asymptotiques sur la complexité d'algorithmes de manipulation;
— ensuite pour éclairer le comportement probabiliste de ces algorithmes.

1. INTRODUCTION

Many tables searching methods have been analysed, mainly in order to obtain their asymptotic complexity. Among them, let us mention: the Interpolation Search, the Interpolation-then-Sequential Search, the Fast-Search, the p-Search.

Tables sorting methods often deal with permutations such as Shellsort.

Complexity results are often obtained by delicate and advanced techniques such as martingales, combinatorial arguments, information theory, summation formulas, ... Very often *ad-hoc* methods have to be devised.

In this preliminary paper, we intend to show that many complexity results can easily be deduced from the properties of a well-know continuous Markov Process: *the Brownian Motion.*

The approach provides a new and usually simpler analysis tool.

(*) Received in April 1982, revised in January 1983.

(1) Université libre de Bruxelles, Laboratoire d'Informatique théorique CP212, boulevard du Triomphe, 1050 Bruxelles, Belgique.

R.A.I.R.O. Informatique théorique/Theoretical Informatics, 0399-0540/1983/365/$ 5.00

The Brownian Motion (also called Wiener process, *see* § 2) has many interesting properties and a lot of other processes (among them the standard random walk) converge, in a suitable sense, to it.

Using these convergences, we are able to obtain new results, or to prove easily old results, on asymptotic complexity of sorted tables manipulations.

We also get a simpler proof of a recent result of Sedgewick [14], on inversions in 2-ordered permutation. The main characteristic of the method is that one first proves the convergence of some stochastic paths to a Brownian Motion (or some variant of it) and then uses classical results on this Motion (such as crossing times) to get asymptotic behaviour of important parameters.

Once these asymptotic results are obtained, standard reasoning lead easily to general complexity results: this reasoning is well described in the aftermentioned papers and we shall not always repeat it. We are mainly interested in finding asymptotic properties of basic probabilities and averages.

The paper is organised as follows: § II is a review of Brownian Motion and its main properties, § III deals with Interpolation Search. Theorems 1, 2, 3, 4 are new. Several previous known results can been easily deduced from them. § IV describes how our method can be applied to the *p*-Search. Theorem 5 is well-known, Theorem 6 is new. Using our methods we get a more process-oriented proof of Theorem 5. This proof could perhaps lead to solving an open problem. § V shows how a Brownian correspondence lemma helps to write a shorter proof on an approximation of the average number of inversions in a 2-ordered permutation. This approximation was previously proved by delicate sommation and combinatorial arguments in Sedgewick [14]. § VI concludes the paper and mentions a few possible areas of future research.

2. THE BROWNIAN MOTION

The Brownian Motion, or Wiener Process, $\eta(u)$ $(0 \leqq u < +\infty)$ is a continuous Markov process with transition density defined as follows (*see* Ito and McKean [10] for detailed description):

$$Pr[\eta(v) \in db \mid \eta(u) = a] = \frac{1}{\sqrt{2\pi(v-u)}} \exp\left[\frac{-(b-a)^2}{2(v-u)}\right] db, \qquad \text{for} \quad u \leqq v.$$

The density is clearly Gaussian. A Brownian Bridge $\xi(t)$ $(0 \leqq t \leqq 1)$ is a Brownian Motion starting form 0 at time 0 and *conditioned* to return to 0 at time 1.

The density of $\xi(t)$ is given by:

$$\frac{1}{\sqrt{2\pi t(1-t)}}\exp\left[\frac{-\xi^2}{2t(1-t)}\right]d\xi$$

and its covariance is given by:

$$E[\xi(u)\xi(v)]=u(1-v) \qquad \text{for} \quad u \leqq v$$

(*see*, for instance, Doob [6], Donsker [5], Ito and McKean [10], p. 40).

We also have the equivalence:

$$\xi(u)=(1-u)\eta\left(\frac{u}{1-u}\right),$$

where $\eta(u)$ $(0 \leqq u < +\infty)$ is an ordinary Brownian Motion.

The crossing time of a Brownian Motion $\eta(u)$ and a straight line $p+qu$ is given by the following density (dating back to Chandrasekhar [3], *see* also Cox and Miller [4], p. 221):

$$Pr[\min(v' : \eta(v')=p+qv') \in dv]=\frac{p}{\sqrt{2\pi}v^{3/2}}\exp\left[\frac{-(p+qv)^2}{2v}\right]dv.$$

3. INTERPOLATION SEARCH

3.1 The algorithm

Let us try to find the location of an existing key α in a sorted table of n values, drawn from a uniform [0, 1] distribution. The interpolation search first probes the position $[n\alpha]$, and repeats iteratively the same search rule in case of failure.

It is well known that the asymptotic complexity of this algorithm is $0(\log\log n)$ (all logarithms are base 2).

See for instance Gonnet *et al.* [8, 9], Yao and Yao [15], Perl *et al.* [13].

3.2 The Brownian Motion perspective

Let F_n be the sample distribution function for a sample of n points drawn from a uniform [0, 1] distribution.

We have the weak convergence (in the uniform convergence topology):

$$\sqrt{n}[F_n(t)-t] \quad \Rightarrow \quad \xi(t), \qquad n \to \infty, \qquad t \in [0, 1],$$

where $\xi(t)$ is a Brownian Bridge (Brownian Motion starting from 0 at time 0 and conditioned to return to 0 at time 1).

We can then hope to get asymptotic result on the algorithm by using this correspondence: distributions and crossing times have explicitly known expression for this Markov process.

3.3 Asymptotic results on Interpolation Search

Three new asymptotic results can be obtained on the behaviour, at step k, of the probability $P(k)$ of success, of the searched key $\alpha(k)$ and of the sample dimension $n(k)$.

All the proofs show a common pattern: the correspondence with the Brownian Motion is used for the first step ($k=1$) and one then proceeds by induction.

The asymptotic result of Theorem 3 can be summarized in an easily constructed and interpreted binary tree.

THEOREM 1: *The probability $P(k)$ of a success at step k is asymptotically given by:*

$$P(k) \sim \psi(k) \frac{1}{2\sqrt{\pi}} \frac{1}{\theta^{2-k}},$$

where $\sim$ denotes the asymptotic n-equivalence:

$$\theta = \frac{n}{2}\alpha(1-\alpha),$$

$$\psi(1) = 1,$$

$$\psi(k) = \prod_{j=2}^{k} \left[\Gamma\left(\frac{1}{2} - 2^{-j}\right) \Big/ \sqrt{\pi} \right], \qquad k \geqq 2.$$

Proof: (*a*) The probability of a success at the first consultation ($k=1$) is asymptotically given by:

$$P(1) \sim Pr\left[(F_n(\alpha) - \alpha) \in \left[0, \frac{1}{n}\right] \right] \sim Pr\left[\frac{\xi(\alpha)}{\sqrt{n}} \in \left[0, \frac{1}{n}\right] \right]$$

$$= Pr\left[\xi(\alpha) \in \left[0, \frac{1}{\sqrt{n}}\right] \right] \sim \frac{1}{\sqrt{2\pi\alpha(1-\alpha)}} \frac{1}{\sqrt{n}} = \frac{1}{2\sqrt{\pi}\sqrt{\theta}};$$

(*b*) In case of failure, the density of the value u of the key observed in position $[n\alpha]$ is asymptotically given by:

$$\sim Pr\left[\min\left(u' : u' + \frac{\xi(u')}{\sqrt{u'}} = \alpha\right) \in du\right]$$

$$= Pr[\min(u' : \xi(u') = \sqrt{n}(\alpha - u')) \in du]$$

$$= Pr\left[\min\left(u' : (1-u')\eta\left(\frac{u'}{1-u'}\right) = \sqrt{n}(\alpha - u')\right) \in du\right],$$

where η is a classical Brownian Motion.

Let $u'/(1-u') = v'$. We get:

$$Pr[\min(v' : \eta(v') = \alpha\sqrt{n} + \sqrt{n}(\alpha - 1)v') \in dv]$$

$$= \frac{\alpha\sqrt{n}}{\sqrt{2\pi}v^{3/2}} \exp\left[\frac{-n(\alpha + (\alpha-1)v)^2}{2v}\right] dv,$$

by the result mentioned in § II.

Let $u = \alpha + y$. We get, after a few manipulations, the following asymptotic density for y:

$$\frac{\sqrt{n}}{\sqrt{2\pi\alpha(1-\alpha)}} \exp\left[\frac{-ny^2}{2\alpha(1-\alpha)}\right] dy, \tag{1}$$

which shows that y is a Normal variable with standard deviation:

$$\frac{\sqrt{\alpha(1-\alpha)}}{\sqrt{n}}. \tag{2}$$

One can show that this asymptotic distribution is still correct for $\alpha = n^{-(1-\varepsilon)}$, $0 < \varepsilon < 1$.

(*c*) Let us now proceed by induction.

Let the probability of a success at step $(k-1)$ be given by:

$$P(k-1) \sim \frac{1}{2\sqrt{\pi}} \psi(k-1)\theta^{-2^{-(k-1)}}.$$

In case of failure at the first step, two possibilities can arise: $y > 0$ or $y < 0$.

We are then led to a table of n_1 keys, where we look for the position of key α_1, with:

(i) $y > 0$:

$$n_1 = \alpha n,$$
$$\alpha_1 = \alpha/(\alpha+y).$$

(ii) $y < 0$:

$$n_1 = (1-\alpha)n,$$
$$\alpha_1 = -y/(1-\alpha-y).$$

The probability of success at step k is asymptotically given by:

$$P(k) \sim \frac{\psi(k-1)}{2\sqrt{\pi}} \times \int_{-\infty}^{+\infty} \frac{\sqrt{n}}{\sqrt{2\pi\alpha(1-\alpha)}} \exp\left[\frac{-ny^2}{2\alpha(1-\alpha)}\right]\left[\frac{1}{2}n_1\alpha_1(1-\alpha_1)\right]^{-2^{-(k-1)}} dy$$

$$= \frac{\psi(k-1)}{2\sqrt{\pi}}\left\{\int_0^{\infty} \frac{\sqrt{n}}{\sqrt{2\pi\alpha(1-\alpha)}} \exp\left[-\frac{ny^2}{2\alpha(1-\alpha)}\right]\left[\frac{2(\alpha+y)^2}{\alpha n \alpha y}\right]^{2^{-(k-1)}} dy \right.$$

$$\left. + \int_{-\infty}^{0} \frac{\sqrt{n}}{\sqrt{2\pi\alpha(1-\alpha)}} \exp\left[-\frac{ny^2}{2\alpha(1-\alpha)}\right]\left[\frac{2(1-\alpha-y)^2}{(-y)(1-\alpha)^2 n}\right]^{2^{-(k-1)}} dy\right\}$$

$$\sim \frac{\psi(k-1)}{2\sqrt{\pi}} 2 \int_0^{\infty} \frac{\sqrt{n}}{\sqrt{2\pi\alpha(1-\alpha)}} \exp\left[\frac{-ny^2}{2\alpha(1-\alpha)}\right]\left[\frac{2}{yn}\right]^{2^{-(k-1)}} dy.$$

Let $z = ny^2/(2\alpha(1-\alpha))$. We get after transformation:

$$P(k) \sim \frac{\psi(k-1)}{2\sqrt{\pi}} \int_0^{\infty} \frac{e^{-z}}{\sqrt{\pi z}}\left[\frac{2}{\sqrt{n2\alpha(1-\alpha)}\sqrt{z}}\right]^{2^{-(k-1)}} dz$$

$$= \frac{\psi(k-1)}{2\sqrt{\pi}} \int_0^{\infty} \frac{e^{-z}}{\sqrt{\pi} z^{(1/2+2^{-k})}}\left[\frac{2}{n\alpha(1-\alpha)}\right]^{2^{-k}} dz$$

$$= \frac{\psi(k-1)}{2\sqrt{\pi}}\left[\frac{\Gamma(1/2-2^{-k})}{\sqrt{\pi}}\right]\theta^{-2^{-k}} = \frac{\psi(k)}{2\sqrt{\pi}}\theta^{-2^{-k}}$$

∇

THEOREM 2: *Let $\alpha(k)$ and $n(k)$ be the searched key and the sample dimension at step k.*

Let:

$$\theta(k)=\frac{n(k)}{2}\alpha(k)[1-\alpha(k)],$$

with:

$$\theta(1)\equiv\theta=\frac{n}{2}\alpha(1-\alpha).$$

We have asymptotically:

$$E[\theta(k)]\sim\varphi(k)\theta^{2^{-(k-1)}},$$

with:

$$\varphi(1)=1,$$

$$\varphi(k)=\prod_{j=2}^{k}[\Gamma(1/2+2^{-(j-1)})/\sqrt{\pi}],\qquad k\geqq 2.$$

Proof: Proceeding as in Theorem 1, we get for $k\geqq 2$:

$$E[\theta(k)]\sim\varphi(k-1)\int_0^\infty\frac{e^{-z}}{\sqrt{\pi z}}\left[\sqrt{\frac{n}{2}\alpha(1-\alpha)}\sqrt{z}\right]^{2^{-(k-2)}}dz$$

$$=\varphi(k-1)\left[\frac{\Gamma(1/2+2^{-(k-1)})}{\sqrt{\pi}}\right]\theta^{2^{-(k-1)}}=\varphi(k)\theta^{2^{-(k-1)}}.\quad\nabla$$

REMARK: We know (Bateman [1], p. 6) that:

$$\Gamma(v)=2^{2v(1-2^{-n})-n}\Gamma(2^{-n}v)\prod_{m=1}^{n}[\pi^{-1/2}\Gamma(1/2+2^{-m}v)].\tag{3}$$

We can then write ψ and φ in a compact form:

(*a*) let $v=-1/2$ in (3). We get:

$$\Gamma(-1/2)=2^{-(1-2^{-k})-k}\Gamma(-2^{-(k+1)})\prod_{m=1}^{k}[\pi^{-1/2}\Gamma(1/2-2^{-(m+1)})]$$

and:

$$\psi(k+1)=\frac{-2\sqrt{\pi}2^{(1-2^{-k})+k}}{\Gamma(-2^{-(k+1)})}=\frac{2\sqrt{\pi}2^{-2^{-k}}}{\Gamma(1-2^{-(k+1)})}$$

and:

$$\psi(k) \underset{k \to \infty}{\uparrow} 2\sqrt{\pi}, \tag{4}$$

(*b*) let $v=1$ in (3). We get:

$$\Gamma(1)=2^{2(1-2^{-k})-k}\Gamma(2^{-k}) \prod_{m=1}^{k} [\pi^{-1/2}\Gamma(1/2+2^{-m})]$$

and:

$$\varphi(k+1)=\frac{2^{-2+2^{-k+1}+k}}{\Gamma(2^{-k})}=\frac{2^{-2+2^{-k+1}}}{\Gamma(1+2^{-k})}$$

and:

$$\varphi(k) \underset{k \to \infty}{\downarrow} 1/4. \tag{5}$$

THEOREM 3: *Let* $\alpha(k)$ *and* $n(k)$ *defined as in Theorem* 2. *Let* $\alpha^*(k)=\min[\alpha(k), (1-\alpha(k))]$ *and* $D(k)=\alpha^*(k)n(k)$. *We get asymptotically:*

$$E[D(k)] \sim 2\varphi(k)\theta^{2^{-(k-1)}}.$$

Proof: Proceeding as before, we get:

$$E[D(2)] \sim \int_0^\infty \frac{e^{-z}}{\sqrt{\pi z}}[\sqrt{2n\alpha(1-\alpha)}\sqrt{z}]dz=\frac{2}{\sqrt{\pi}}\sqrt{\theta}=2\varphi(2)\sqrt{\theta}.$$

Induction is then straightforward. ∇

It is interesting to analyse the asymptotic behaviour of $\alpha(k)$, $n(k)$ and $y(k)$ during the successive steps.

Let $\alpha^*=\min(\alpha, (1-\alpha))$. We get:

$$y(k) \sim 0\left(\sqrt{\frac{\alpha^*(k)}{n(k)}}\right) \qquad \textit{see (1)}$$

and:

$$\alpha^*(k+1) \sim 0\left(\frac{y(k)}{\alpha^*(k)}\right), \qquad n(k+1) \sim 0(\alpha^*(k)n(k)),$$

or:

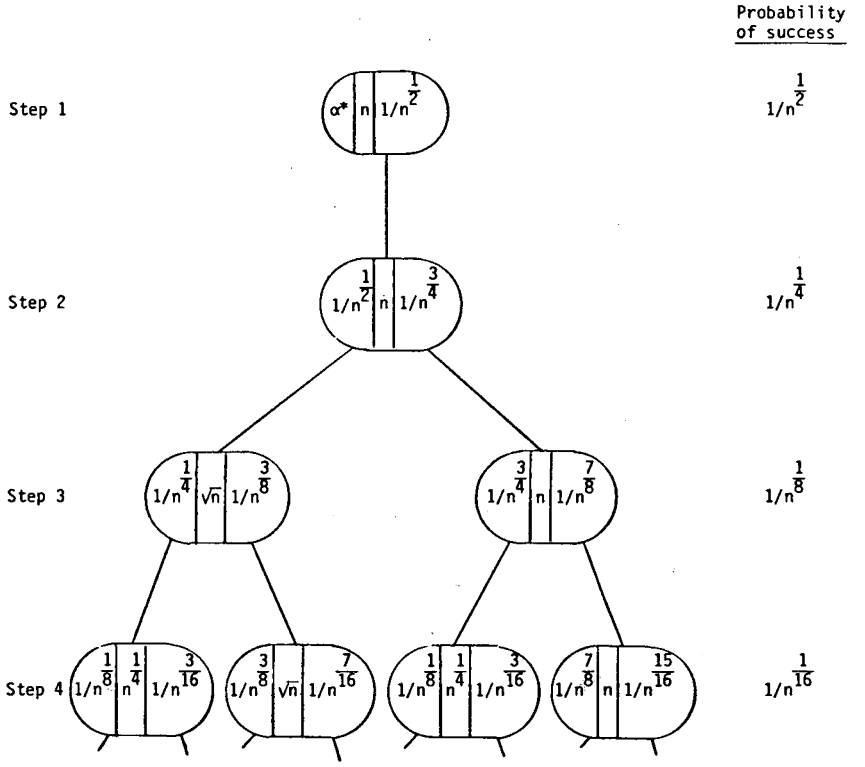
$$\alpha^*(k+1) \sim 0(y(k)), \qquad n(k+1) \sim 0(n(k))$$

[*see* part (*c*) of Theorem's 1 proof)].

The probability $P(k)$ of success at step k is given by:

$$P(k) \sim 0\left(\frac{1}{\sqrt{\alpha^*(k)\, n(k)}}\right).$$

All possible configurations of the triplet $\boxed{\alpha^*(k) \mid n(k) \mid y(k)}$ can be summarized in an asymptotic binary tree as follows.



The asymptotic behaviour of the probability of success conforms to Theorem 1.

3.4 Connection with previous results

Some asymptotic lemmas used by several authors in complexity analysis are easily deduced from our theorems. Let us mention the following examples.

(*a*) From (2) we asymptotically obtain $Pr[y \geqq k\sqrt{\alpha/n}] \leqq 1/k^2$ (by Tchebycheff inequality) and letting:

$$h = \alpha n,$$

$$\varepsilon(h) = 1/\sqrt{\log h},$$

$$k = \frac{h^{\varepsilon(h)}}{8},$$

we get:

$$Pr\left[y \geqq \sqrt{\frac{\alpha}{n}}\frac{h^{\varepsilon(h)}}{8}\right] \leqq \frac{64}{h^{2\varepsilon(h)}} < \frac{128}{h^{2\varepsilon(h)}},$$

which corresponds to Lemma 3 of Yao and Yao [15]; this lemma is crucial in their proof of the $0(\log\log n)$ complexity of Interpolation Search.

(*b*) from Theorem 1, we get $P(1) \sim 1/(2\sqrt{\pi\theta})$ which is identical to Gonnet [8] (3.5.9.) and Gonnet *et al.* [9] (3.11) (to compare equations, let us mention that θ_G used in Gonnet [8 and 9]: $\theta_G \equiv n\alpha(1-\alpha) = 2\theta$ in this paper).

(*c*) From Theorem 2, we get $E(\theta(2)) \sim \sqrt{\theta}/\sqrt{\pi}$ which is identical to Gonnet [8] (3.4.36) and Gonnet *et al.* [9] (3.25).

But, from the equivalent result:

$$E[\pi\theta(2)] \sim \sqrt{\pi\theta},$$

Gonnet gets (by Jensen inequality):

$$E[\pi\theta(k)] \leqq [\pi\theta]^{2^{-(k-1)}},$$

which is less precise than the asymptotic result (for large k) we get from Theorem 2 and (5):

$$E[\theta(k)] \sim \frac{1}{4}\theta^{2^{-(k-1)}}.$$

(*d*) From Theorem 1 and (4) we get (for large k):

$$P(k) \sim \theta^{-2^{-k}},$$

which is more precise than the result of Gonnet [8] (3.5.12) and Gonnet *et al.* [9] (3.35):

$$P[k] \geqq 1/2(\pi\theta)^{-2^{-k}}.$$

(*e*) Using the same method as in Theorem 3, we get:

$$E[\alpha^*(2)\, n(2)]^2 \sim n\alpha(1-\alpha),$$

which corresponds to inequality (7) of Perl *et al.* [13].

(*f*) For large k, we get from (5):

$$E[D(k)] \sim \frac{1}{2}\theta^{2^{-(k-1)}}$$

but:

$$\frac{1}{2}\theta^{2^{-(k-1)}} < n^{2^{-(k-1)}}.$$

We get then, asymptotically, inequality (8) of Perl *et al.* [13].

3.5 Extension to other interpolation algorithms

3.5.1. *Interpolation-then-sequential search*

In case of failure after the first interpolation search, we run systematically through the remaining table.

The following theorem gives a new result on the mean number of consultations $Ps(\alpha)$ in case of first failure.

THEOREM 4: *The mean number of consultations $Ps(\alpha)$ in case of first failure is asymptotically given by:*

$$Ps(\alpha) \sim \left[\frac{2n\alpha(1-\alpha)}{\pi}\right]^{1/2}.$$

Proof: Proceeding as before, we get:

$$Ps(\alpha) \sim 2\int_0^\infty \frac{\sqrt{n}}{\sqrt{2\pi\alpha(1-\alpha)}} \exp\left[\frac{-ny^2}{2\alpha(1-\alpha)}\right] yn\, dy$$

$$= \int_0^\infty \frac{e^{-z}}{\sqrt{\pi}} [2n\alpha(1-\alpha)]^{1/2}\, dz = \left[\frac{2n\alpha(1-\alpha)}{\pi}\right]^{1/2}. \quad \nabla$$

COROLLARY: *The average number consultations on* [0, 1] *is given by* $\sqrt{n\pi/32}$.

Proof: Immediate by integration of $P(s, \alpha)$ on [0, 1].

The last result is identical to Gonnet [8] (3.8.18).

3.5.2. *Fast Search*

The aim of this generalization of Interpolation Search is to prevent deterioration arising from non-uniform distribution. See Burton *et al.* [2], Lewis *et al.* [12].

The authors propose two modifications:

(i) at the beginning, the table is consulted only from position $\sqrt{n}$ to position $n-\sqrt{n}$ (the gap is initialized to $\sqrt{n}$);

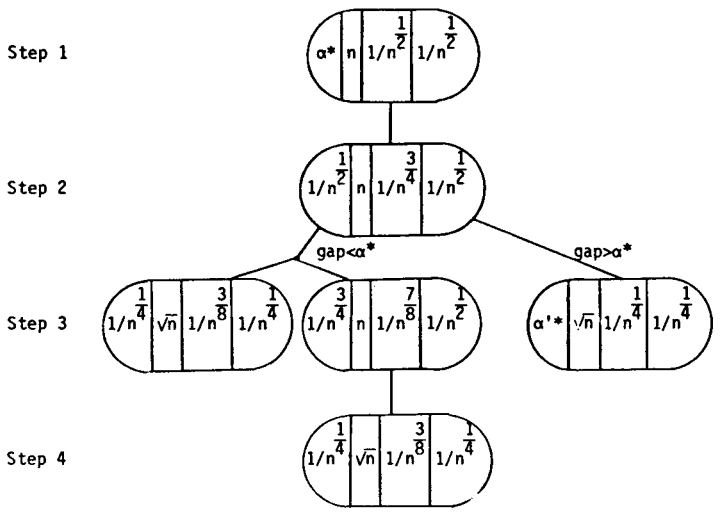
(ii) in case of failure and if the searched key lies in the largest of the two subintervals, the gap is doubled.

These modifications are used at each subtable consultation (*see* details in the two mentioned papers).

In [12], the authors prove that the length of a consulted subtable is reduced to $\sqrt{n}$ in constant expected consultations number.

Taking again the binary tree of § 3, we add a new parameter "gap" to each triplet α^*, n, y.

As the gap, in [0, 1] scale, is $1/\sqrt{n}$, we get the following asymptotic tree:



We check immediately that, asymptotically, in 4 steps maximum, we reduce the length from n to $0(\sqrt{n})$.

4. P-SEARCH

4.1. The algorithm

This technique probes (iteratively) the position $[pn]$ in the sorted table where p is a fixed parameter (Binary Search and Fibonaccian Search are particular cases of p-search, see Knuth [11], § 6.2.1).

It would be quite useful to get the average number $E(\alpha, n)$ of consultations necessary to find key α. This is an open problem. Nevertheless, we know an approximation of the α-average of this quantity:

$$E(n) = \int_0^1 E(\alpha, n)\, d\alpha.$$

Indeed, it is proved (Knuth [11], § 6.2.1, Ex. 20) that:

THEOREM 5: *An approximation of $E(n)$ is asymptotically given by:*

$$E(n) \sim \log_b n \quad \text{where} \quad b = p^p q^q.$$

Proof: The short proof of Knuth goes as follows:

$$E(n) = 1 + p\,E(pn) + q\,E(qn), \qquad n > 1,$$

with:

$$E(1) = 0,$$

which gives immediately $E(n) = \log_b n$. ∇

4.2. Asymptotic probabilities

The probabilities of success are characterized by the following new result:

THEOREM 6. — *Let us denote the 2^k consultation intervals, at step k, resulting from k successive subdivisions, by $d_{i,k}$, $i = 1 \ldots 2^k$ [these intervals result from developing $(p+q)^k$]. Let the position of the point c_i be given by:*

$$c_i = \sum_{l=0}^{i-1} d_{l,k} + p d_{i,k} \quad (\text{by convention } d_{0,k} \equiv 0), \qquad (i = 1 \ldots 2^k).$$

The probability of success at step $k+1$ (as function of α) *is asymptotically given by 2^k Gaussian curves, centered at 2^k points c_i. The Gaussian function centered in c_i is given by:*

$$\frac{1}{\sqrt{2\pi npq}\sqrt{d_{i,k}}}\exp\left[\frac{-n(\alpha-c_i)^2}{(2pq)d_{i,k}}\right].$$

The area under each curve is given by $1/n$.

Proof:

(i) Proceeding as in §2, we see that, asymptotically, the probability of success at the first step: $P(1, \alpha, n)$ is given by (letting $q=1-p$):

$$P(1,\alpha,n)\sim\frac{1}{\sqrt{2\pi\alpha(1-\alpha)}}\exp\left[-\frac{n(p-\alpha)^2}{2\alpha(1-\alpha)}\right]\frac{1}{\sqrt{n}}$$

$$\sim\frac{1}{\sqrt{2\pi pq}}\exp\left[-\frac{n(p-\alpha)^2}{2pq}\right]\frac{1}{\sqrt{n}}$$

$$\text{as}\quad\frac{\alpha-p}{\sqrt{2(1-\alpha)}}\ \text{must be}\ 0\left(\frac{1}{\sqrt{n}}\right).$$

(ii) In case of failure at step 1, the asymptotic density of the value u of the key found in position $[pn]$ is asymptotically given by (letting $u=p+y$):

$$\frac{\sqrt{n}}{\sqrt{2\pi pq}}\exp\left[-\frac{ny^2}{2pq}\right]dy \qquad [\textit{see}\ (1)]$$

if $\alpha<p$, the probability of success at step 2 is given by:

$$\int_{-\infty}^{\alpha-p}\frac{\sqrt{n}}{\sqrt{2\pi pq}}\exp\left[-\frac{ny^2}{2pq}\right]P\left(1,\frac{\alpha-p-y}{q-y},nq\right)dy$$

$$+\int_{\alpha-p}^{\infty}\frac{\sqrt{n}}{\sqrt{2\pi pq}}\exp\left[-\frac{ny^2}{2pq}\right]P\left(1,\frac{\alpha}{p+y},np\right)dy.$$

We see immediately that the first integral is asymptotically (exponentially) negligible and we get by our usual transformation:

$$P(2, \alpha, n) \sim \int_0^{\infty} \frac{e^{-z}}{\sqrt{\pi z}} \frac{1}{\sqrt{2\pi pq}} \exp\left[\frac{-np(p-(\alpha/p))^2}{2pq}\right] \frac{1}{\sqrt{np}} dz$$

$$= \frac{1}{\sqrt{2\pi pq}} \frac{1}{\sqrt{np}} \exp\left[\frac{-n(\alpha-p^2)^2}{(2pq)p}\right].$$

Similarly, if $\alpha > p$, we get:

$$\frac{1}{\sqrt{2\pi pq}} \frac{1}{\sqrt{nq}} \exp\left[\frac{-n(\alpha-p-pq)^2}{(2pq)q}\right].$$

The probability of success at step 2 is thus asymptotically given (as function of α) by two Gaussian-like functions, centered at p^2 and $p+pq$, with standard-deviation:

$$\sqrt{\frac{pq}{n}}\sqrt{p} \quad \text{and} \quad \sqrt{\frac{pq}{n}}\sqrt{q},$$

with height:

$$\frac{1}{\sqrt{2\pi npq}} \frac{1}{\sqrt{p}} \quad \text{and} \quad \frac{1}{\sqrt{2\pi npq}} \frac{1}{\sqrt{q}}$$

respectively.

The area under each curve is $1/n$.

(iii) Induction or k is straightforward. ∇

4.3. Another approach to Theorem 5

Although we have not yet succeded in computing asymptotics for $E(\alpha, n)$, we can get another proof of Theorem 5.

Indeed, in first approximation, the discovery of α happens when step k corresponds to an interval $d_{i,k}$ of length $1/n$ (containing in the average 1 key) and covering α.

A first approximation of $E(n)$ is thus given by:

$$\sum_0^{\infty} k \sum_{\{i \; p^i q^{k-i} = 1/n\}} \binom{k}{i} p^i q^{k-i}. \tag{6}$$

To compute this approximation, we observe that:

(*a*) for large n, k is also large;

(*b*) k and i can be replaced by real-valued variables κ and ι;

(*c*) the binomial distribution may be replaced by a normal:

$$\mathcal{N}(kp, \sqrt{kpq});$$

(*d*) the set $\{i: p^i q^{k-i} = 1/n\}$ must be replaced by a set of intervals:

$$[\iota^*(\kappa) - \beta_1(\kappa, n) \leqq \iota \leqq \iota^*(\kappa) + \beta_2(\kappa, n)],$$

containing the value $\iota^*(\kappa)$ defined by:

$$p^{\iota^*} q^{\kappa - \iota^*} = 1/n \tag{7}$$

and such that the set of all these intervals partition *exactly* [0, 1].

These intervals are defined later.

Finally (6) becomes:

$$E(n) \sim \int_0^\infty \kappa \int_{\iota^* - \beta_1(\kappa, n)}^{\iota^* + \beta_2(\kappa, n)} \exp\left[-\frac{(\iota - \kappa p)^2}{2\kappa pq} \right] \frac{1}{\sqrt{2\pi\kappa pq}} \, d\iota \, d\kappa; \tag{8}$$

(*e*) the maximum on κ of the Gaussian is realized in κ^* defined by:

$$\kappa^* p = \iota^*,$$

i. e.:

$$p^{\kappa^* p} q^{\kappa^* q} = 1/n \qquad \text{or} \qquad \kappa^* = \log_b n. \tag{9}$$

Letting:

$$\xi = \iota^* - \kappa p, \qquad \gamma_1(\eta, n) = \beta_1(\kappa, n),$$
$$\eta = \kappa - \kappa^*, \qquad \gamma_2(\eta, n) = \beta_2(\kappa, n)$$

(8) becomes:

$$\sim \int_{-\infty}^{+\infty} (\kappa^* + \eta) \frac{[\gamma_2(\eta, n) + \gamma_1(\eta, n)]}{\sqrt{2\pi\kappa^* pq}} \exp\left[-\frac{\xi^2}{2\kappa^* pq} \right] d\eta; \tag{10}$$

(*f*) (7) can also be written as:

$$p^{\iota^* - \kappa p + \kappa p - \kappa^* p + \kappa^* p}, \qquad q^{\kappa - \iota^* - \kappa q + \kappa q - \kappa^* q + \kappa^* q} = 1/n,$$

or, with (9):

$$p^{\xi+p\eta}q^{-\xi+q\eta}=1,$$

i. e.:

$$\xi=\delta\eta \qquad \text{with} \quad \delta=\frac{-p\log p-q\log q}{\log p-\log q}$$

(*g*) γ_1 and γ_2 can be computed as follows. Let $p<q$.

Let k be fixed and let $d_{j,k}$ be one interval $p^j q^{k-j}$.

To each interval of this kind corresponds, at step $k+1$, 2 intervals: $pd_{j,k}$ and $qd_{j,k}$.

In order to partition exactly [0, 1], it is necessary that the *largest unused interval* in the sum:

$$\sum_{j=\iota-\gamma_1}^{\iota+\gamma_2-1}\binom{k}{j}p^j q^{k-j}, \tag{11}$$

correspond (by p-subdivision) to the *smallest interval* at step $k+1$ which gives:

$$p\left(\frac{q}{p}\right)^{\gamma_1+1}\frac{1}{n}=\left(\frac{p}{q}\right)^{\gamma_2-1}\frac{1}{n},$$

or:

$$\gamma_1+\gamma_2=\frac{\log p}{\log p-\log q}. \tag{12}$$

But we see then the first term of (11) also gives, by q-subdivision, an interval $1/n\,(p/q)^{\gamma_2-1}$ at step $k+1$.

To prevent counting these intervals twice, it is necessary to substract q from (12), which finally gives:

$$\gamma_1+\gamma_2=\frac{p\log p+q\log q}{\log p-\log q}\equiv-\delta.$$

The case $p>q$ is similarly treated.

(*h*) We finally get from (10):

$$E(n)\sim\int_{-\infty}^{+\infty}(\kappa^*+\eta)\frac{1}{\sqrt{2\pi\kappa^* pq}}\exp\left[-\frac{(\delta\eta)^2}{2\kappa^* pq}\right]\delta\,d\eta=\kappa^*. \quad \nabla$$

Of course our (lengthy) proof is rather complex, especially compared with Knuth's proof! However:

— it sheds some light on the detailed asymptotic behaviour of success probabilities;

— we believe that it is the right approach to try the computation of $E(\alpha, n)$.

5. SPEED-UP OF A PROOF: INVERSIONS IN TWO-ORDERED PERMUTATION

5.1. Two-ordered permutations

A practical improvement to classical insertion sort, called Shellsort, makes several passes through a file, each time sorting h independent subfiles of elements spaced by h. In attempting to analyse the simplest version in which h takes value 2, we need the average number $A(n)$ of inversions in a 2-ordered permutation of the $2n$ values $\{1 \ldots 2n\}$ i.e. a permutation consisting of two interleaved sorted permutations. See for instance Knuth [11], § 5.2.1, Ex. 15.

5.2. Approximation of $A(n)$

THEOREM 7. — *An approximation of $A(n)$ is given by:*

$$A(n) \sim \sqrt{\pi n^3}/4.$$

This theorem is proved in Sedgewick [14], p. 162-166, by delicate summation and combinatorial arguments. Sedgewick mentions that a shorter proof is available through a combinatorial generating function argument: using Knuth's correspondence to paths in a lattice diagram it is possible to show that the generating function:

$$B(w, z) = \sum_{\text{all 2 ordered perms } P} w^{|P|} z^{\text{inv}(P)}$$

[where $\text{inv}(P)$ is the number of inversions in P] satisfies:

$$B_w(w, z)\big|_{z=1} = \frac{w}{(1-4w)^2}.$$

However, this derivation not only involves an indirect argument using the generating function for particular types of paths in the lattice but also some complicated manipulations with derivatives of these generating functions.

5.3. Another proof of Theorem 7

Let us first prove a correspondence lemma.

Denote by $\varphi(i, i+j)$ the probability that position i of the odd part of the permutation contains value $i+j$. Letting $k=i+j$, $\varphi(i, i+j)$ is given by (*see* Sedgewick [14], p. 163):

$$\binom{k-1}{k-i}\binom{2n-k}{n-k+i}\Big/\binom{2n}{n}.$$

Let the path $u(i, k)$, corresponding to a given partition, be defined as follows: $u(i)=k$ if position i of the odd part of the permutation contains value k.

LEMMA: *Path u is asymptotically equivalent to a Brownian Bridge.*

Proof: Let:

$$x=\frac{k}{2n}, \qquad y=\frac{i}{n},$$

so that:

$$\frac{2n-k}{2n}=1-x \qquad \text{and} \qquad \frac{k-i}{n}=2x-y.$$

By Stirling's formula we get:

$$\varphi(y, x) \sim \frac{x^k(1-x)^{2n-k}\sqrt{x(1-x)}}{\sqrt{\pi n}y^i(1-y)^{n-i}\sqrt{y(1-y)}(2x-y)^{k-i}(1-2x+y)^{n-k+i}\sqrt{(2x-y)(1-2x+y)}}.$$

Letting now:

$$x=y+\frac{z}{\sqrt{2n}} \quad \text{so that} \quad 2x-y=x+\frac{z}{\sqrt{2n}}=y+\frac{2z}{\sqrt{2n}},$$

and using the well-know formula:

$$\left(1+\frac{z}{n}\right)^n=e^z\left[1-\frac{z^2}{2n}+0\left(\frac{1}{n^2}\right)\right],$$

We get after a few manipulations the asymptotic density of z:

$$\frac{\exp-[z^2/2y(1-y)]}{\sqrt{2\pi y(1-y)}}dz.$$

The covariance of z is checked by the same method. ∇

Theorem 7 is now easily proved as follows.

Proof:

$$A(n)=\sum_{i=1}^{n}\sum_{j=0}^{n}|i-j-1|\varphi(i,\ i+j) \quad (\textit{see}\ \text{Sedgewick [14], p. 163}),$$

but:

$$|i-j|=|2i-k|=|y-x|2n=\sqrt{2n}|z|.$$

Asymptotically, we then get:

$$A(n)\sim\sqrt{2n}\int_0^1 n dy\int_{-\infty}^{+\infty}\frac{\exp[-z^2/2y(1-y)]}{\sqrt{2\pi y(1-y)}}|z|dz=\frac{n^{3/2}\sqrt{\pi}}{4}. \quad \nabla$$

5.4. Some comments

It is clear again that finding a correspondence between some paths and a Brownian processes is a useful key in getting asymptotic formulas. Indeed the main problem is usually to find and prove the correspondence.

6. CONCLUSION

Using the Brownian Motion and Gaussian-like functions we can get easily old and new asymptotic results on sorted tables manipulation (i. e. search and sorting algorithms). A few of our results are sometimes less refined than what we can (in some cases) get by delicate probability and combinatorial techniques.

But, in addition to putting all these results in an unified framework, the present approach seems to shed more light on the asymptotic behaviour of some algorithms. The basic idea is first to show a correspondence between some stochastic paths and Brownian Motion or some variant of it (through some weak convergence defined in a suitable sense).

We can then use all the collection of known results on Brownian Motion properties to get asymptotic formulas for basic probabilities or averages.

We intend to pursue this approach on some open problems such as:

– computation of the average number $E(\alpha, n)$ of consultations necessary to find, with p-Search, key α in a sorted table of n keys;

– asymptotic analysis of valued-path (*see* for instance Flajolet [7], Chap. IV).

ACKNOWLEDGEMENTS

We are indebted to a referee for helpful comments leading to a substantial improvement of this paper's presentation.

REFERENCES

1. H. BATEMAN, *Higher Transcendal Functions*, Vol. 1, McGraw-Hill, 1953.
2. F. W. BURTON and G. N. LEWIS, *A Robust Variation of Interpolation Search*, Information Processing Letters, Vol. 10, No. 4 and 5, 1980, pp. 198-201.
3. S. CHANDRASEKHAR, *Stochastic Problems in Physics and Astronomy*, Review of Modern Physics, Vol. 15, 1943, pp. 57-59.
4. D. R. COX and H. D. MILLER, *The Theory of Stochastic Processes*, Chapman and Hall, 1980.
5. M. D. DONSKER, *Justification and Extension of Doob's Heuristic Approach to the Kolmogorov-Smirnov Theorems*, Annals of Mathematical Statistics, 1952, pp. 277-281.
6. J. L. DOOB, *Heuristic Approach to the Kolmogorov-Smirnov Theorems*, Annals of Mathematical and Statistics, Vol. 20, 1949, pp. 393-403.
7. P. FLAJOLET, *Analyse d'algorithmes de manipulation d'arbres et de fichiers*, Cahiers du BURO, 1981, pp. 34-35.
8. G. H. GONNET, *Interpolation and Interpolation-Hash Searching*, Research Report CS-77-02, University of Waterloo, 1977.
9. G. H. GONNET, D. R. LAWRENCE and J. A. GEORGE, *An Algorithmic and Complexity Analysis of Interpolation Search*, Acta Informatica, Vol. 13, 1980, pp. 39-52.
10. K. ITO and Jr. H. P. MCKEAN, *Diffusion Processes and their Sample Paths*, Springer-Verlag, 1974.
11. D. E. KNUTH, *The Art of Computer Programming*, Vol. 3, Addison-Wesley, 1973.
12. G. N. LEWIS, N. J. BOYNTON and F. W. BURTON, *Expected Complexity of Fast Search with Uniformly Distributed Data*, Information Processing Letters, Vol. 13, No. 1, 1981, pp. 4-7.
13. Y. PERL, A. ITAI and H. AVNI, *Interpolation Search. A Log Log N Search*, Communications of the ACM, Vol. 21, No. 7, 1978, pp. 550-553.
14. R. SEDGEWICK, *Mathematical Analysis of Combinatorial Algorithms in Probability and Computer Science*, G. LATOUCHE and G. LOUCHARD, Ed., Academic Press (to appear).
15. A. C. YAO and F. F. YAO, *The Complexity of Searching an Ordered Random Table*, Proceedings of the 17th Annual Symposium on Foundations of Computer Science, 1976, pp. 173-177.