2007 Conference on Analysis of Algorithms, AofA 07 DMTCS proc. **AH**, 2007, 237–246

Sorting using complete subintervals and the maximum number of runs in a randomly evolving sequence: Extended abstract.

Svante Janson

Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden
svante.janson@math.uu.se *and* http://www.math.uu.se/∼svante/

received 9th February 2007, revised 18th May 2007,

We study the space requirements of a sorting algorithm where only items that at the end will be adjacent are kept together. This is equivalent to the following combinatorial problem: Consider a string of fixed length n that starts as a string of 0's, and then evolves by changing each 0 to 1, with the n changes done in random order. What is the maximal number of runs of 1's?

We give asymptotic results for the distribution and mean. It turns out that, as in many problems involving a maximum, the maximum is asymptotically normal, with fluctuations of order $n^{1/2}$, and to the first order well approximated by the number of runs at the instance when the expectation is maximized, in this case when half the elements have changed to 1; there is also a second order term of order $n^{1/3}$.

We also treat some variations, including priority queues and sock-sorting.

Keywords: sorting algorithm, runs, priority queues, sock-sorting, evolution of random strings, Brownian motion

1 Introduction

Gunnar af Hällström [1] considered, as indicated at the end of his paper, the following algorithm for sorting an unordered pile of student exams in alphabetic order. (It is said that he used this procedure himself.)

The exams are taken one by one from the input. The first exam is put in a new pile. For each following exam (x, say), if the name on it is immediately preceding the name on an exam y at the top of one of the piles, the new exam x is put on top of y. (The professor knows the names of all the students, and can thus see that there are no names between x and y.) Similarly, if the name on x is immediately succeeding the name on an exam z at the bottom of a pile, x is put under z. If both cases apply, with y on top of one pile and z at the bottom of another, the two piles are merged with x inserted between z and y. Finally, if there is no pile matching x in one of these ways, x is put in a new pile.

The algorithm thus maintains a list of sorted piles, each being an interval without gaps of the set of exams. At the end, there is a single sorted pile.

The problem is the space requirement of this algorithm; more precisely, the maximum number of sorted piles during the execution. The input is assumed to be in random order, so this is a random variable, and we are interested in its mean and distribution.

af Hällström [1] gave the following mathematical reformulation. Consider a deck of n cards numbered $1, \dots, n$ in random order, and a sequence of n places with the same numbers in order. Take the cards one by one and put them at their respective places. When we have placed m cards, $0 \le m \le n$, we see $X_{n,m}$ "islands", i.e. uninterrupted blocks of cards. What is $X_n^* := \max_m X_{n,m}$?

Alternatively, in the language of parking cars: n cars park, one by one, on n available places along a street; each car parks at a random free place. What is the maximum number of uninterrupted blocks of cars during the process?

Let, for $n \ge 1$, $0 \le m \le n$ and $1 \le k \le n$, the indicator $I_{n,m}(k)$ be 1 if the item (exam or card) with number k is one of the m first in the input, and 0 otherwise. Thus, $X_{n,m}$ is the number of runs of 1's in the random sequence $I_{n,m}(1), \dots, I_{n,m}(n)$ of $n-m$ 0's and m 1's. Note that each random sequence $(I_{n,m}(k))_{k=1}^n$ is uniformly distributed over all $\binom{n}{m}$ possibilities; moreover, for each $m < n$ we obtain $(I_{n,m+1}(k))_{k=1}^n$ from $(I_{n,m}(k))_{k=1}^n$ by changing a single randomly chosen 0 to 1, this random choice being uniform among the $n-m$ 0's, and independent of the previous history. (In other words, the order the digits are changed is given by a random permutation of $[n]$.) Hence, X_n^* is the maximum number of runs seen during this process.

This quantity was also studied, under the name *clustering* of a permutation, by Flajolet [8], who gave formulas for the generating function of the number of permutations with a given clustering; these formulas, derived from continued fractions, involve Laguerre polynomials. It seems difficult, however, to obtain asymptotic results of the type given in this paper from Flajolet's exact formulas.

It is easy to see that $\mathbb{E}\, X_{n,m} = m(n-m+1)/n$, see (3.1); it follows that the maximum of $\mathbb{E}\, X_{n,m}$ for a given n is attained for $m = \lceil n/2 \rceil$, and that $\mathbb{E}\, X_{n,\lceil n/2 \rceil} > n/4$. Since obviously

$$\mathbb{E}\, X_n^* = \mathbb{E} \max_m X_{n,m} \ge \max_m \mathbb{E}\, X_{n,m}, \tag{1.1}$$

this yields $\mathbb{E}\, X_n^* > n/4$ as observed by af Hällström [1]. Moreover, he observed that $\mathbb{E}\, X_n^*$ is subadditive, and thus the limit

$$\gamma := \lim_{n\to\infty} \mathbb{E}\, X_n^*/n$$

exists and equals $\inf_n \mathbb{E}\, X_n^*/n$; he further showed that $1/4 \le \gamma \le 1/3$, where the lower bound comes from (1.1). Based on simulations with $n = 13$ and $n = 52$, af Hällström [1] concluded that γ seems to be very close to or equal to 1/4. We will show that, indeed, $\gamma = 1/4$. We also show that the distribution of X_n^* is asymptotically normal, with a variance of order n.

Theorem 1.1 *As $n \to \infty$,*

$$n^{-1/2}\big(X_n^* - n/4\big) \xrightarrow{\mathrm{d}} N(0, 1/16),$$

with convergence of all moments. In particular,

$$\mathbb{E}\, X_n^* = n/4 + o(n^{1/2}),$$
$$\operatorname{Var} X_n^* = n/16 + o(n).$$

This theorem says that to the first order, the maximum number of piles (runs) X_n^* behaves like the number $X_{n,m}$ with $m = \lceil n/2 \rceil$. A more refined analysis shows that the difference $X_n^* - X_{n,\lceil n/2 \rceil}$ is of order $n^{1/3}$. Let $B(t)$, $-\infty < t < \infty$, be a standard two-sided Brownian motion.

Theorem 1.2 *As $n \to \infty$,*

$$n^{-1/3}\bigl(X_n^* - X_{n,\lceil n/2\rceil}\bigr) \xrightarrow{\mathrm{d}} \tfrac{1}{2}V,$$

where the random variable V is defined by $V := \max_t\bigl(B(t) - t^2/2\bigr)$, and

$$\mathbb{E}\,X_n^* = \mathbb{E}\,X_{n,\lceil n/2\rceil} + \tfrac{1}{2}\,\mathbb{E}\,V n^{1/3} + o(n^{1/3}) = \tfrac{1}{4}n + \tfrac{1}{2}\,\mathbb{E}\,V n^{1/3} + o(n^{1/3}).$$

The random variable V is studied by Barbour [2], Daniels and Skyrme [7] and Groeneboom [11]. Note that $0 < V < \infty$ a.s. We have, see [7] (using `Maple` to improve the numerical values in [2; 3; 7; 6]), with Ai the Airy function,

$$\mathbb{E}\,V = -\frac{2^{-1/3}}{2\pi}\int_{-\infty}^{\infty}\frac{\mathrm{i}y\,\mathrm{d}y}{\mathrm{Ai}(\mathrm{i}y)^2} \approx 0.996193.$$

The numerical values $X_{13}^* \approx 4.22$ and $X_{52}^* \approx 14.66$ found experimentally by af Hällström [1] differ from $n/4$ by about 18% and 10% less than the correction term $\frac{1}{2}\,\mathbb{E}\,V n^{1/3}$ in Theorem 1.2, which is a reasonable agreement for such rather small n.

We prove these theorem by studying asymptotics of the entire (random) process $(X_{n,m})_{m=0}^n$. The natural time here is m/n, so we take $m = \lfloor nt\rfloor$ for $0 \le t \le 1$ and consider the process $X_{n,\lfloor nt\rfloor}$ with a continuous parameter $t \in [0,1]$. The following theorem shows that this process asymptotically is Gaussian. (Recall that a Gaussian process is a random function $Z(t)$ such that the value for a fixed t has a normal distribution, and the values $Z(t_1), \dots, Z(t_k)$ for any fixed finite sequence $t_1, \dots, t_k$ have a joint normal distribution. Recall also that the distribution of a Gaussian process is determined by the means and covariances of $Z(t)$. The space $D[0,1]$ below is the standard space of right-continuous functions with left-hand limits on [0,1], equipped with the Skorohod topology, see [4].)

Theorem 1.3 *As $n \to \infty$, in the space $D[0,1]$ of functions on $[0,1]$,*

$$n^{-1/2}\bigl(X_{n,\lfloor nt\rfloor} - nt(1-t)\bigr) \xrightarrow{\mathrm{d}} Z(t),$$

where Z is a continuous Gaussian process on $[0,1]$ with mean $\mathbb{E}\,Z(t) = 0$ and covariances

$$\mathbb{E}\bigl(Z(s)Z(t)\bigr) = s^2(1-t)^2, \qquad 0 \le s \le t \le 1. \tag{1.2}$$

The behaviour of X_n^* shown in Theorems 1.1 and 1.2, with an asymptotic normal distribution with a mean of order n and random fluctuations of order $n^{1/2}$, and with a second order term for the mean of order $n^{1/3}$, is common for this type of random variables defined as the maximum of some randomly evolving process. For various examples, both combinatorial and others, and general results see for example Daniels [5; 6], Daniels and Skyrme [7], Barbour [2; 3] and Louchard, Kenyon and Schott [20]. Indeed, paraphrasing the explanations in these papers, in many such problems, the first order asymptotic of a random process $X_n(t)$ (after suitable scaling) is a deterministic function $f(t)$, say, defined on a compact interval I (typically scaled to be $[0,1]$ as here). Hence the first order asymptotic of the maximum of the process is just the maximum of this function f. Moreover, it is often natural to expect that the random fluctuations around this function $f(t)$ asymptotically form a Gaussian process $G(t)$; this is then a second order term of smaller order as in our Theorem 1.3. If we assume that f is continuous on I and has a unique maximum at a point $t_0 \in I$, then the maximum of the process $X_n(t)$ is attained close to

t_0, so the first order approximation of the maximum is the constant $f(t_0) = \max_t f(t)$, while the next approximation is just $X_n(t_0)$, giving a normal limit law as in our Theorem 1.1. The Gaussian fluctuations in this limit have mean 0, so in order to find the next term for the mean $\mathbb{E}\, X_n^*$, we study more closely the difference $\max_t X_n(t) - X_n(t_0)$ by studying the difference $X_n(t) - X_n(t_0)$ close to t_0. Assuming that t_0 is an interior point of I and that f is twice differentiable at t_0 with $f''(t_0) \neq 0$, we can locally at t_0 approximate f by a parabola and $G(t) - G(t_0)$ by a two-sided Brownian motion (with some scaling), and thus $\max_t X_n(t) - X_n(t_0)$ is approximated by a scaling constant times the variable V above, see Barbour [2] and, in our case, Theorem 4.1 below. In the typical case where the mean of $X_n(t)$ is of order n and the Gaussian fluctuations are of order $n^{1/2}$, it is easily seen that the correct scaling gives, as in Theorem 1.2 above, a correction to $\mathbb{E}\, X_n^*$ of order $n^{1/3}$, see [2; 5; 6]. (Sketch of the argument: $X_n(t)$ is approximated by $nf(t) + \sqrt{n}G(t)$, so consider the maximum of the latter. At a point $t_0 + x$ with $|x|$ small, $n(f(t) - f(t_0))$ is negative and of the order nx^2, while $n^{1/2}(G(t) - G(t_0))$ is like a constant times $n^{1/2}B(x)$ and thus of order $n^{1/2}x^{1/2}$. Typically, the maximum occurs when these two terms are of the same order (with opposite signs); thus $nx^2 = \Theta(n^{1/2}x^{1/2})$ or $x = \Theta(n^{-1/3})$ and the maximum is $\Theta(nx^2) = \Theta(n^{1/2}x^{1/2}) = \Theta(n^{1/3})$.)

The method used in the present paper is a simple adaption of the method used in [12] and [13] to study the number of subgraphs of a given isomorphism type in a random graph. These papers study the random graphs $G(n,p)$ and $G(n,m)$ that can be constructed by random deletion of edges in the complete graph K_n (with the deletions being independent for $G(n,p)$ and such that a fixed number of edges are deleted for $G(n,m)$). The method applies more generally to random graphs constructed by random edge deletions in these ways from any fixed initial graph F_n. The problem treated in this paper can be regarded as an instance of this when the initial graph is the path P_n with n edges.

Our method applies also to other problems. One example is given by *priority queues*, where Louchard [18] and Louchard, Kenyon and Schott [20] have proved asymptotic results very similar to the Theorems 1.1–1.3 above, see Section 5 for details. In particular, they found the same asymptotic covariance (1.2) except for a normalizing constant. An equivalent problem is the *width* of involutions, another case treated by Flajolet [8]; he gives formulas for the generating function, derived from continued fractions, which in this case involve Hermite polynomials. See also Flajolet, Françon and Vuillemin [9], Flajolet, Puech and Vuillemin [10] and Lagarias, Odlyzko, and Zagier [16]. Again, we do not see how to obtain asymptotic results from these exact combinatorial results.

Priority queues can be defined as follows. Suppose that n items are to be temporarily stored (or processed); let item i arrive at time A_i and be deleted at time D_i. We assume that the $2n$ times A_i and D_i are distinct; thus they can be arranged in a sequence of the $2n$ events A_i and D_i, with A_i coming before D_i for each i. We assume further, as our probabilistic model, that all $(2n)!/2^n$ such sequences are equally probable. Ignoring the labels, we can equivalently consider sequences of n A and n D (or n $+$ and n $-$), where each A is paired with a D coming later; there is a 1–1 correspondence between such sequences and pairings of $1, \dots, 2n$ into n pairs, and there are $(2n-1)!! = (2n)!/(2^n n!)$ such sequences (with pairings), again taken with equal probability.

Let, for $m = 0, \dots, 2n$, $Y_{n,m}$ be the number of items stored after m of these events, i.e. the number of A's minus the number of D's among the m first events, and let $Y_n^* := \max_{0 \le m \le 2n} Y_{n,m}$. The sequence $(Y_{n,m})_0^{2n}$ is a Dyck path, but note that its distribution is not uniform; for a given Dyck path (or a given sequence of A and D without labels), the number of ways to pair a given D with a preceding A, i.e. the number of ways to choose which item to delete, equals the current number of items stored before this deletion. Thus, the weight of the Dyck path equals the product of these numbers $\prod_{m: Y_{n,m+1} < Y_{n,m}} Y_{n,m}$.

Alternatively, which better explains the name priority queue, we can keep the stored items in a list showing the order in which they eventually will be deleted; then there is only one choice for each deletion but each new item can be inserted in $Y+1$ ways if there are Y items stored before the insertion, and thus $Y+1$ after it; hence the weight can also be written as $\prod_{m:Y_{n,m}>Y_{n,m-1}} Y_{n,m}$. (It is easy to see directly that the two products are equal.)

An equivalent example is *sock-sorting*, studied by Li and Pritchard [17] and Steinsaltz [22]. Suppose that we have $2n$ socks; the socks form n pairs with the two socks in each pair identical but different from all others. All socks are mixed and we pick them in random order. If the picked sock is from a pair that we have not yet seen, it is put on a bench; on the other hand, if we already have picked the other sock in the pair, that sock is taken from the bench, paired with its twin, and put away in permanent storage. What is the maximum number of socks on the bench? It is easily seen that this is equivalent to a priority queue.

Our method applies to priority queues and socks too, and we obtain the same asymptotic results as for $X_{n,m}$ and X_n^*. (Note that there is no exact correspondence for finite n, since the natural sample spaces have $n!$ elements for $X_{n,m}$ but $(2n-1)!!$ elements for $Y_{n,m}$.) Again, we can regard the problem as an instance of subgraph counts for randomly deleting edges from a given initial graph F_n; in this case taking F_n to be a multigraph consisting of n double edges.

Another example is a model suggested by Van Wyk and Vitter [24] as a model for *hashing with lazy deletion*, and further studied by Louchard [19] and Louchard, Kenyon and Schott [20]. In this model, n items arrive and are deleted as above, but now the arrival and deletion times A_i and D_i are random numbers, with the n pairs (A_i, D_i) mutually independent and each pair distributed as $\bigl(\min(T_i,\tilde{T}_i),\ \max(T_i,\tilde{T}_i)\bigr)$, where T_i and $\tilde{T}_i$ are independent random variables uniformly distributed on [0,1]. We let $Y_n(t)$ be the number of items present at time t, and again we are especially interested in its maximum $\max_t Y_n(t)$. Again, the asymptotic results for the maximum found by Louchard, Kenyon and Schott [20] are the same as in our Theorems 1.1 and 1.2, except for a constant factor, while the asymptotic result for the process $Y_n(t)$ found by Louchard [19] differs somewhat from the one in Theorem 1.3, it corresponds instead to the one in Theorem 2.1 below; see Section 5. Indeed, as explained by Kenyon and Vitter [15], this model can be seen as a priority queue with randomized times for insertions and deletions, which explains why the results for the maximum are the same as for priority queues.

Proofs are given in the journal version of this paper [14].

2 Randomizing time

We will use the standard method of randomizing the time. More precisely, we let $T_1, \dots, T_n$ be independent random variables, each uniformly distributed on $(0,1)$. We interpret T_k as the time item k arrives, and note that a.s. there are no ties. We define

$$I(t;k) = \mathbf{1}[T_k \le t],$$

i.e., $I(t;k)=1$ if item k has arrived by time t. We further define $N_n(t)$ as the number of items that have arrived at time t, and $X_n(t)$ as the number of runs of 1's at time t, i.e.,

$$N_n(t)=\sum_{k=1}^{n} I(t;k), \tag{2.1}$$

$$X_n(t)=I(t;1)+\sum_{k=1}^{n-1}\bigl(1-I(t;k)\bigr)I(t;k+1) \tag{2.2}$$

$$=N_n(t)-\sum_{k=1}^{n-1} I(t;k)I(t;k+1). \tag{2.3}$$

Clearly, the items arrive in random order, so the process remains the same except that the insertions occur at the random times $\{T_1,\dots,T_n\}$. In particular,

$$X_n^*=\max_{0\le t\le 1} X_n(t).$$

For the process $X_n(t)$, we have the following analogue of Theorem 1.3; note that the (co)variances differ.

Theorem 2.1 *As $n\to\infty$, in $D[0,1]$,*

$$n^{-1/2}\bigl(X_n(t)-nt(1-t)\bigr)\xrightarrow{\mathrm{d}} Z(t),$$

where Z is a continuous Gaussian process on $[0,1]$ with mean $\mathbb{E}\,Z(t)=0$ and covariances, for $0\le s\le t\le 1$,

$$\begin{aligned}\mathbb{E}\bigl(Z(s)Z(t)\bigr)&=s(1-2s)(1-t)(1-2t)+s^2(1-t)^2\\&=s(1-t)(1-s-2t+3st).\end{aligned}$$

The importance of this randomization is that the variables $I(t;k)$, $k=1,\dots,n$, are independent. Thus, $X_n(t)$ is the number of runs of 1 in a sequence of *independent* 0's and 1's, each with the distribution $\mathrm{Be}(t)$. Furthermore, the number of items sorted at time t is $N_n(t)\sim\mathrm{Bi}(n,t)$.

Define further, for $0\le t\le 1$, the centralized variables

$$I'(t;k):=I(t;k)-\mathbb{E}\,I(t;k)=I(t;k)-t$$

and the sums

$$S_{n,1}(t):=\sum_{k=1}^{n} I'(t;k)=N_n(t)-\mathbb{E}\,N_n(t)=N_n(t)-nt, \tag{2.4}$$

$$S_{n,2}(t):=\sum_{k=1}^{n-1} I'(t;k)I'(t;k+1). \tag{2.5}$$

Thus $\mathbb{E}\, S_{n,1}(t) = \mathbb{E}\, S_{n,2}(t) = 0$ for all $t \in [0,1]$. We have

$$\begin{aligned} N_n(t) &= \sum_{k=1}^{n} \bigl(I'(t;k) + t\bigr) = S_{n,1}(t) + nt, \\ \sum_{k=1}^{n-1} I(t;k)I(t;k+1) &= \sum_{k=1}^{n-1} \bigl(I'(t;k) + t\bigr)\bigl(I'(t;k+1) + t\bigr) \\ &= S_{n,2}(t) + t\bigl(2S_{n,1}(t) - I'(t;1) - I'(t;n)\bigr) + (n-1)t^2, \end{aligned}$$

and thus from (2.3) the representation

$$\begin{aligned} X_n(t) &= n(t - t^2) + t^2 + (1-2t)S_{n,1}(t) - S_{n,2}(t) + tI'(t;1) + tI'(t;n) \\ &= nt(1-t) + (1-2t)S_{n,1}(t) - S_{n,2}(t) + R_n(t), \end{aligned} \tag{2.6}$$

where $R_n(t) := t^2 + tI'(t;1) + tI'(t;n)$ and thus $|R_n(t)| \le 3$.

Note that for any fixed t, the variables $I'(t;k)$ are independent and have means 0; hence the terms in the sums in (2.4) and (2.5) have means and all covariances 0. (They are thus orthogonal in L^2.) It follows immediately that

$$\begin{aligned} \operatorname{Var}\,(S_{n,1}(t)) &= n\,\mathbb{E}\bigl(I'(t;1)\bigr)^2 = n\operatorname{Var}\bigl(I(t;1)\bigr) = nt(1-t), \\ \operatorname{Var}\,(S_{n,2}(t)) &= (n-1)t^2(1-t)^2, \\ \operatorname{Cov}\bigl(S_{n,1}(t), S_{n,2}(t)\bigr) &= 0. \end{aligned}$$

3 Exact results

It is easy to find the exact distribution of $X_{n,m}$ for given n and m, see for example Stevens [23] or Mood [21]. The mean and variance can easily be computed:

$$\mathbb{E}\, X_{n,m} = \frac{m(n-m+1)}{n}, \tag{3.1}$$

$$\operatorname{Var} X_{n,m} = \frac{m(m-1)(n-m)(n-m+1)}{n^2(n-1)}. \tag{3.2}$$

If we instead randomize the insertion times as in Section 2 and consider the process at a fixed time t, we find from (2.2) and the independence of $I(t;k)$ for $k = 1, \dots, n$,

$$\begin{aligned} \mathbb{E}\, X_n(t) &= t + \sum_{k=1}^{n-1} (1-t)t = nt(1-t) + t^2. \\ \operatorname{Var} X_n(t) &= nt(1-t)(1-3t+3t^2) + t^2(1-t)(3-5t). \end{aligned}$$

To find the exact distribution of X_n^* seems much more complicated, but exact values of $\mathbb{P}(X_n^* = h)$ can be calculated from the expressions for generating functions given by Flajolet [8]. See also af Hällström [1] for some exact values for small n.

4 The method of proof

The main idea is to first prove joint convergence of the normalized processes $n^{-1/2}S_{n,1}(t)$ and $n^{-1/2}S_{n,2}(t)$ to two independent continuous Gaussian processes on $[0,1]$. This is done, as the corresponding results in [12] and [13], using martingale theory, in particular a continuous time martingale limit theorem. Then, the representation (2.6) yields the asymptotic distribution of the process $X_n(t)$ in Theorem 2.1. Furthermore, using the corresponding (and simpler) limit result for $N_n(t)$, it is possible to "derandomize" the time and obtain Theorem 1.3. Finally, a closer study of the process close to the point $t = 1/2$ where $\mathbb{E}\, X_n(t)$ has its maximum yield the results for the maximum X_n^*, shows that $X_n(t)$ there, after rescaling, converges to a two-sided Brownian motion with parabolic drift.

Theorem 4.1 *As $n \to \infty$, in $D(-\infty,\infty)$,*

$$n^{-1/3}\big(X_n(\tfrac{1}{2} + xn^{-1/3}) - X_n(\tfrac{1}{2})\big) \xrightarrow{\mathrm{d}} 2^{-1/2}B(x) - x^2,$$

where B is a Brownian motion on $(-\infty,\infty)$.

5 Priority queues, sock-sorting and lazy hashing

As said above, priority queues are equivalent to sock-sorting. Furthermore, randomization of the times, as in Section 2 in these problems thus gives exactly the model for lazy hashing defined in Section 1, as found by Kenyon and Vitter [15]. In particular, $\max_t Y_n(t) \overset{\mathrm{d}}{=} Y_n^*$.

Our method then yields the following results, corresponding to our results above for $X_{n,m}$ and $X_n(t)$.

Theorem 5.1 *As $n \to \infty$, in $D[0,1]$,*

$$n^{-1/2}\big(Y_n(t) - 2nt(1-t)\big) \xrightarrow{\mathrm{d}} Z(t) := (1-2t)Z_1(t) - 2Z_2(t),$$

where Z is a continuous Gaussian process on $[0,1]$ with mean $\mathbb{E}\, Z(t) = 0$ and covariances, for $0 \le s \le t \le 1$,

$$\begin{aligned}\mathbb{E}\big(Z(s)Z(t)\big) &= 2s(1-2s)(1-t)(1-2t) + 4s^2(1-t)^2 \\ &= 2s(1-t) - 4s(1-s)t(1-t).\end{aligned}$$

Theorem 5.2 *As $n \to \infty$, in $D[0,1]$,*

$$n^{-1/2}\big(Y_{n,\lfloor 2nt \rfloor} - 2nt(1-t)\big) \xrightarrow{\mathrm{d}} Z(t) := -2Z_2(t),$$

where Z is a continuous Gaussian process on $[0,1]$ with mean $\mathbb{E}\, Z(t) = 0$ and covariances

$$\mathbb{E}\big(Z(s)Z(t)\big) = 4s^2(1-t)^2, \qquad 0 \le s \le t \le 1.$$

Theorem 5.3 *As $n \to \infty$, in $D(-\infty,\infty)$,*

$$n^{-1/3}\big(Y_n(\tfrac{1}{2} + xn^{-1/3}) - Y_n(\tfrac{1}{2})\big) \xrightarrow{\mathrm{d}} 2^{1/2}B(x) - 2x^2,$$

where B is a Brownian motion on $(-\infty,\infty)$.

Theorem 5.4 *As $n \to \infty$,*

$$n^{-1/2}\big(Y_n^* - n/2\big) \xrightarrow{\mathrm{d}} N(0, 1/4),$$

with convergence of all moments. In particular,

$$\mathbb{E}\, Y_n^* = n/2 + o(n^{1/2}),$$
$$\operatorname{Var} Y_n^* = n/4 + o(n).$$

Theorem 5.5 *As $n \to \infty$,*

$$n^{-1/3}\big(Y_n^* - Y_{n,n}\big) \xrightarrow{\mathrm{d}} V,$$

where the random variable V is as in Theorem 1.2, and

$$\mathbb{E}\, Y_n^* = \mathbb{E}\, Y_{n,n} + \mathbb{E}\, V n^{1/3} + o(n^{1/3}) = \tfrac{1}{2}n + \mathbb{E}\, V n^{1/3} + o(n^{1/3}).$$

Theorem 5.1 is given by Louchard [19], Theorem 5.2 by Louchard [18] (with a deterministic change of time, making the problem equivalent to a queueing problem), and Theorems 5.4 and 5.5 by Louchard, Kenyon and Schott [20] (with different proofs).

Note that the limits in Theorem 5.2 and Theorem 1.3 are the same, except for a normalization factor. (Unlike Theorem 2.1 and Theorem 5.1, where the variances of the limits are $t(1-t)(1-3t+3t^2)$ and $2t(1-t)(1-2t+2t^2)$.)

Acknowledgement 1 *I thank Göran Högnäs for telling me about this problem and providing me with the reference [1]. I also thank Persi Diaconis, Anders Martin-Löf and Guy Louchard for interesting discussions.*

References

[1] G. af Hällström, Ein lineares Inselproblem der kombinatorischen Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.* **123** (1952), 9 pp.

[2] A.D. Barbour, A note on the maximum size of a closed epidemic. *J. Roy. Statist. Soc. Ser. B* **37** (1975), no. 3, 459–460.

[3] A.D. Barbour, Brownian motion and a sharply curved boundary. *Adv. Appl. Probab.* **13** (1981), no. 4, 736–750.

[4] P. Billingsley, *Convergence of Probability Measures*. Wiley, New York, 1968.

[5] H. E. Daniels, The maximum size of a closed epidemic. *Adv. Appl. Probab.* **6** (1974), 607–621.

[6] H. E. Daniels, The maximum of a Gaussian process whose mean path has a maximum, with an application to the strength of bundles of fibres. *Adv. Appl. Probab.* **21** (1989), no. 2, 315–333.

[7] H. E. Daniels and T. H. R. Skyrme, The maximum of a random walk whose mean path has a maximum. *Adv. Appl. Probab.* **17** (1985), no. 1, 85–99.

[8] P. Flajolet, Combinatorial aspects of continued fractions. *Discrete Math.* **32** (1980), no. 2, 125–161.

[9] P. Flajolet, J. Françon & J. Vuillemin, Sequence of operations analysis for dynamic data structures. *J. Algorithms* **1** (1980), no. 2, 111–141.

[10] P. Flajolet, C. Puech & J. Vuillemin, The analysis of simple list structures. *Inform. Sci.* **38** (1986), no. 2, 121–146.

[11] P. Groeneboom, Brownian motion with a parabolic drift and Airy functions. *Probab. Theory Related Fields* **81** (1989), no. 1, 79–109.

[12] S. Janson, A functional limit theorem for random graphs with applications to subgraph count statistics, *Random Struct. Alg.* **1** (1990), 15–37.

[13] S. Janson, *Orthogonal Decompositions and Functional Limit Theorems for Random Graph Statistics*. Mem. Amer. Math. Soc., vol. 111, no. 534, American Mathematical Society, Providence, R.I., 1994.

[14] S. Janson, Sorting using complete subintervals and the maximum number of runs in a randomly evolving sequence. `http://arXiv.org/math.PR/0701288`

[15] C. M. Kenyon & J. S. Vitter, Maximum queue size and hashing with lazy deletion. *Algorithmica* **6** (1991), no. 4, 597–619.

[16] J. C. Lagarias, A. M. Odlyzko, D. B. Zagier, On the capacity of disjointly shared networks. *Comput. Networks ISDN Systems* **10** (1985), no. 5, 275–285.

[17] W. V. Li & G. Pritchard, A central limit theorem for the sock-sorting problem. *High dimensional probability (Oberwolfach, 1996)*, Progr. Probab., 43, Birkhäuser, Basel, 1998, 245–248.

[18] G. Louchard, Random walks, Gaussian processes and list structures. *Theoret. Comput. Sci.* **53** (1987), no. 1, 99–124.

[19] G. Louchard, Large finite population queueing systems. I. The infinite server model. *Comm. Statist. Stochastic Models* **4** (1988), no. 3, 473–505.

[20] G. Louchard, C. Kenyon & R. Schott, Data structures' maxima. *SIAM J. Comput.* **26** (1997), no. 4, 1006–1042.

[21] A. M. Mood, The distribution theory of runs. *Ann. Math. Statistics* **11** (1940), 367–392.

[22] D. Steinsaltz, Random time changes for sock-sorting and other stochastic process limit theorems. *Electron. J. Probab.* **4** (1999), no. 14, 25 pp.

[23] W. L. Stevens, Distribution of groups in a sequence of alternatives. *Ann. Eugenics* **IX** (1939), 10–17.

[24] C. J. Van Wyk & J. S. Vitter, The complexity of hashing with lazy deletion. *Algorithmica* **1** (1986), no. 1, 17–29.