*2007 Conference on Analysis of Algorithms, AofA 07* DMTCS proc. **AH**, 2007, 451–462

# *Distributional asymptotics in the analysis of algorithms: Periodicities and discretization*

Rudolf Grübel

*Institut für Mathematische Stochastik, Leibniz Universität Hannover, Postfach 60 09, D-30060 Hannover, Germany*

*received 19 Feb 2007*, *revised 18 Mar 2007*, *accepted . . .*.

It is well known that many distributions that arise in the analysis of algorithms have an asymptotically fluctuating behaviour in the sense that we do not have 'full' convergence, but only convergence along suitable subsequences as the size of the input to the algorithm tends to infinity. We are interested in constructions that display such behaviour via an ordinarily convergent background process in the sense that the periodicities arise from this process by deterministic transformations, typically involving discretization as a decisive step. This leads to structural representations of the resulting family of limit distributions along subsequences, which in turn may give access to their properties, such as the tail behaviour (unsuccessful search in digital search trees) or the dependence on parameters of the algorithm (success probability in a selection algorithm).

**Keywords:** counting processes, digital search trees, geometric distribution, loser election, Markov chains, renewal processes, Sukhatme-Rényi representation, von Neumann addition

## 1 Introduction

Suppose that we have an algorithm where some quantity $X_n$ of interest, typically the required number of operations of a particular type, depends on a parameter $n$, typically the size of the input to the algorithm. We assume that $X_n$ is random, either because the input is random, or because of the random decisions made while carrying out the procedure (or both). It is then a familiar phenomenon that such quantities, possibly after a suitable normalization, show a fluctuating behaviour as $n \to \infty$ in the sense that we have distributional convergence of $X_n$ only along suitable subsequences $(n_k)_{k\in\mathbb{N}}$, but not along the full sequence $n = 1, 2, 3, \ldots$. The set of distributions is then relatively compact but there is a whole family $Q_\eta$ of limit distributions indexed by $\eta$ in some set $M$ that represents the accumulation points. In a typical case we may have convergence of the distribution $\mathcal{L}(X_{n_k})$ of $X_{n_k}$ if the fractional part $\{\log n_k\}$ of the logarithm of $n_k$ converges to a fixed value $\eta \in M = [0, 1]$ as $k \to \infty$, i.e. we approximate $\mathcal{L}(X_n)$ by $Q_{\{\log n\}}$. Note that $t \mapsto Q_{\{\log t\}}$ is a logarithmically periodic function. Often these periodic fluctuations are very small and hence difficult to detect numerically.

In the present paper we look at various situations where this phenomenon has been observed over the years; particular cases that we will consider below are the number of iterations required by von Neumann addition, the number of comparisons needed to insert an item in a digital search tree, and the number of losers in a simple election algorithm. For these examples, there is an interesting interplay between discretization and renewal type arguments involving counting processes.

Many researchers have observed and analyzed the asymptotic oscillations that appear in the analysis of many algorithms; a survey of the classical analytic approach, mainly dealing with the expectation and the variance associated with $Q_\eta$, is given in (Pro04). Here we hope to contribute to the understanding of such periodicity phenomena by 'resolving' them with the help of probabilistic constructions, for example in the form of a background sequence $(Y_n)_{n\in\mathbb{N}}$ of random variables that converge in the ordinary sense, where the construction should be such that (the distribution of) $X_n$ is (the distribution of) a deterministic function of $Y_n$. It seems to belong to the folklore of the subject that this can often be done, and that the deterministic function may be as simple as rounding to the next integer. As a consequence of such a construction we may be able to display the limit distributions $Q_\eta$, $\eta \in [0,1]$, as transformations $T_\eta$ of a single distribution $Q$, the limit distribution of the $Y_n$'s as $n \to \infty$. This in turn can be used to read off structural properties of the family $\{Q_\eta :\ 0 \le \eta \le 1\}$ from the behaviour of $T_\eta$ as $\eta$ varies. Also, a property of interest for a specific $Q_\eta$ may sometimes be easier to obtain by first investigating $Q$ and then applying $T_\eta$.

In the next section we give a general result that shows that a condition discussed in (Jan06) in connection with representations by shifts and discretization is closely related to a representation in the above sense. In the following three sections we consider the three examples mentioned above. In all three cases a connection to renewal theory, in its classical form or some inhomogeneous variant, turns up. We show that the fluctuations arise from the discretization of the lifetimes (Section 3) or may be a genuine consequence of sampling an integer-valued counting process (Section 4). We then revisit a selection algorithm (Section 5). We show that the simple direct discretization mechanism via rounding is not possible in that example and that a more complicated construction is needed. We obtain a general representation that also incorporates a basic parameter of the algorithm and hence makes it possible to investigate the behaviour of the quantities of interest, such as the probability that the procedure returns a unique value in the case that the parameter tends to 0. In the final section we mention two other classes of problems where asymptotic distributional fluctuations appear, and where a different approach, not based on discretization, seems to be needed.

We write $H_n = \sum_{j=1}^n 1/j$ for the the $n$th harmonic number, $\mathrm{Exp}(\lambda)$ denotes the exponential distribution with parameter $\lambda$, and $\mathrm{Geo}(p)$ is the geometric distribution with parameter (success probability) $p$. Thus, $\mathcal{L}(X) = \mathrm{Exp}(\lambda)$ or simply $X \sim \mathrm{Exp}(\lambda)$ means that

$$P(X \le x) \;=\; 1 - e^{-\lambda x} \quad \text{for all } x \ge 0, \tag{1}$$

whereas, with $q := 1 - p$,

$$P(X = k) \;=\; pq^{k-1} \quad \text{for all } k \in \mathbb{N} \tag{2}$$

if $\mathcal{L}(X) = \mathrm{Geo}(p)$. We further write $X =_{\text{distr}} Y$ if the random variables $X$ and $Y$ have the same distribution, i.e. if $\mathcal{L}(X) = \mathcal{L}(Y)$, and $X_n \to_{\text{distr}} Y$ if the sequence $(X_n)_{n\in\mathbb{N}}$ converges in distribution to $Y$ as $n \to \infty$. For general results and details concerning convergence in distribution we refer the reader to (Bil68).

## 2 A general representation result

Suppose that we have a sequence $(Y_n)_{n\in\mathbb{N}}$ of real random variables and a sequence $(a_n)_{n\in\mathbb{N}}$ of real numbers such that $Y_n - a_n$ converges in distribution to some random variable $Y$. For simplicity we assume that (the distribution function of) $Y$ is continuous. What can be said about the asymptotic distributional behaviour of $(X_n)_{n\in\mathbb{N}}$, where $X_n := \lceil Y_n \rceil$ for all $n \in \mathbb{N}$? Let $M$ be the set of limit points of the (bounded) sequence $(\{a_n\})_{n\in\mathbb{N}}$ of fractional parts of the shift values in the original convergence statement. If $(n_k)_{k\in\mathbb{N}}$ is such that $\{a_{n_k}\} \to \eta \in M$ as $k \to \infty$ then, by Slutsky's lemma,

$$Y_{n_k} - a_{n_k} + \{a_{n_k}\} \ \to_{\text{distr}} \ Y + \eta. \tag{3}$$

From our assumption on the distribution of $Y$ we obtain that the limit distribution assigns probability 0 to the set of points where $x \mapsto \lceil x \rceil$ is not continuous. Hence we can apply the continuous mapping theorem, and (3) yields

$$X_{n_k} - \lfloor a_{n_k} \rfloor = \lceil Y_{n_k} - a_{n_k} + \{a_{n_k}\} \rceil \ \to_{\text{distr}} \ \lceil Y + \eta \rceil. \tag{4}$$

This shows that rounding may destroy the full convergence, but that we still have convergence along subsequences. From general considerations it is clear that, first, a family of shifted integer-valued random variables can only converge in distribution if the shifts are integers or at least if their fractional parts are eventually all the same, and second, since the family $\{\mathcal{L}(X_n - \lfloor a_n \rfloor) : n \in \mathbb{N}\}$ is tight, that any subsequence $(n_k)_{k\in\mathbb{N}}$ has a subsubsequence $(n_{k_j})_{j\in\mathbb{N}}$ such that $X_{n_{k_j}} - \lfloor a_{n_{k_j}} \rfloor$ converges in distribution as $j \to \infty$. In this simple case we have a necessary and sufficient condition on the subsequences for convergence, and further, the set $\{Q_\eta : \eta \in M\}$ of limit distributions arises from a single distribution $Q = \mathcal{L}(Y)$ by shifts and subsequent discretization.

Is there a converse to this construction? For a sequence $(X_n)_{n\in\mathbb{N}}$ of integer-valued random variables (Jan06) considers the following condition related to convergence by centering: For some function $F$ and some sequence $(a_n)_{n\in\mathbb{N}}$ of real numbers,

$$P(X_n \le k_n) \ = \ F(k_n - a_n) + o(1) \quad \text{as } n \to \infty \tag{5}$$

for all sequences $(k_n)_{n\in\mathbb{N}}$ of integers. Under an additional condition on $F$, a condition that seems to be satisfied in most of the examples arising in the analysis of algorithms, we obtain a general result on the representability by shifts and discretization.

**Theorem 1** *If (5) holds with a continuous distribution function $F$, then there exists a probability space with random variables $Y, Y_1, Y_2, \ldots$ such that*

(i) *$Y_n - a_n$ converges almost surely to $Y$ as $n \to \infty$,*

(ii) *$F$ is the distribution function of $Y$ and*

(iii) *$\mathcal{L}(X_n) = \mathcal{L}(\lceil Y_n \rceil)$ for all $n \in \mathbb{N}$.*

*Proof.* We define a family $\tilde{F}_n$, $n \in \mathbb{N}$, of functions $\tilde{F}_n : \mathbb{R} \to [0,1]$ by

$$\tilde{F}_n(k+x) \ := \ F_n(k) \ + \ \big(F(k - a_n + x) - F(k - a_n)\big) \wedge \big(F_n(k+1) - F_n(k)\big),$$

for all $k \in \mathbb{Z}$, $0 \le x < 1$. It is easy to check that these are distribution functions, and that $\mathcal{L}(X_n) = \mathcal{L}(\lceil Y_n \rceil)$ if $Y_n$ is a random variable with distribution function $\tilde{F}_n$. Further, for an arbitrary $y \in \mathbb{R}$, and with

$$k = k_n := \lfloor y + a_n \rfloor, \quad x = x_n := y + a_n - \lfloor y + a_n \rfloor,$$

we obtain on using (5),

$$\begin{aligned} P(Y_n - a_n \le y) &= \tilde{F}_n(y + a_n) \\ &= \tilde{F}_n(k_n + x_n) \\ &= F_n(k_n) + F(\lfloor y + a_n \rfloor - a_n + y + a_n - \lfloor y + a_n \rfloor) - F(k_n - a_n) + o(1) \\ &= F(y) + o(1). \end{aligned}$$

This shows that $Y_n - a_n$ converges in distribution to $Y$ as $n \to \infty$. The existence of a version that has almost sure convergence follows with Skorohod's representation theorem; see e.g. (Bil86), p.343. □

As explained above, as a consequence of this representation the limit distributions $Q_\eta$, $\eta \in M$, can all be obtained from a single (continuous) distribution $Q$ in the sense that $Q_\eta = \mathcal{L}(\lceil Y + \eta \rceil)$ for all $\eta \in M$, if $Y$ is a random variable with distribution $Q$. This has already been noted by (Jan06).

The additional almost sure convergence, which requires a suitable construction, turns out to be useful in various situations. For example, we may be able to detect some martingales in the representation which makes it possible to use the many important results and techniques from that area; this will be taken up in Section 5.

# 3 From geometric maxima to von Neumann addition

The following is the prototypical situation for the procedure in the previous section: Suppose we have a sequence $(Y_n)_{n\in\mathbb{N}}$ of independent random variables, with $Y_n \sim \mathrm{Geo}(p)$ for all $n \in \mathbb{N}$, and let $M_n := \max\{Y_1, \dots, Y_n\}$ be the maximum of the first $n$ of these. Then, as has been observed a long time ago, $M_n + \lceil \log_q n \rceil$ converges in distribution along a subsequence $(n_k)_{k\in\mathbb{N}}$ if $\lim_{k\to\infty}\{\log_q n_k\} = \eta$ for some $\eta \in [0,1]$. Now suppose that $(Z_n)_{n\in\mathbb{N}}$ is another sequence of independent random variables, with $Z_n \sim \mathrm{Exp}(1)$ for all $n \in \mathbb{N}$. It is well known and easy to check that, with

$$c(p) := -\log(1-p), \tag{6}$$

$(\lceil c(p)^{-1} Z_n \rceil)_{n\in\mathbb{N}}$ is equal in distribution to $(Y_n)_{n\in\mathbb{N}}$ and, further, that $\tilde{M}_n - \log n$ with $\tilde{M}_n := \max\{Z_1, \dots, Z_n\}$ converges in distribution to $Q$, the Gumbel distribution, as $n \to \infty$. Since we obviously have $M_n =_{\text{distr}} \lceil c(p)^{-1} \tilde{M}_n \rceil$ for all $n \in \mathbb{N}$, and as $\lceil \log_q n \rceil = -\lfloor c(p)^{-1} \log n \rfloor$, we see that the limit distribution $Q_\eta$ arises as the image of $Q$ under the transformation $T_\eta : \mathbb{R} \to \mathbb{Z}$, $x \mapsto \lceil c(p)^{-1} x + \eta \rceil$, as in Section 2.

This simple example can serve as the basis for an asymptotic distributional analysis of von Neumann addition. This algorithm is presented in (BvN46) (which, incidentally, contains one of the first average case analyses of an algorithm), see also (Sco85). We directly jump to the corresponding random input model and the quantity $X_n$ of interest: Working with some fixed base $b \in \{2, 3, \dots\}$, the input consists of two streams $x_1, x_2, \dots$ and $y_1, y_2, \dots$ of independent random variables that are uniformly distributed on the set $\{0, 1, \dots, b-1\}$ of available ciphers. Whenever an overflow occurs, i.e. if $x_i + y_i \ge b$, a carry bit propagates as long as $x_{i+j} + y_{i+j} = b - 1$, $j = 1, 2, \dots$, and the number of iterations required for input $(x_1, \dots, x_n)$ and $(y_1, \dots, y_n)$ essentially depends on the maximum length of such discrete intervals in the range from 1 to $n$. The length of these carry intervals has a geometric distribution with parameter $(b-1)/2$, but the number of such intervals started up to time $n$ is now a random quantity $N_n$,

$$N_n := \#\{1 \le i \le n : X_i + Y_i \ge b\}.$$

Note that $(N_n)_{n\in\mathbb{N}_0}$ is the counting process for the number of overflows ($N_n$ and the length of the carry intervals are not independent, though). Using this the following result has been obtained in (GR01), where

$$d_{\mathrm{TV}}\big(\mathcal{L}(X),\mathcal{L}(Y)\big) \;=\; \frac{1}{2}\sum_{j\in\mathbb{Z}}\big|P(X=j)-P(Y=j)\big|$$

denotes the total variation distance between the (distributions of the) integer-valued random variables $X$ and $Y$: With $\mathcal{L}(Z)=Q$ and $\kappa_b := \log_b(b-1)-\log_b 2$,

$$\lim_{n\to\infty} d_{\mathrm{TV}}\Big(X_n - \lfloor \log_b n\rfloor,\, \big\lceil (\log b)^{-1}Z + \kappa_b + \{\log_b n\}\big\rceil\Big) \;=\; 0.$$

The fact that we now have a sample of random size $N_n$ rather than of fixed size $n$ as in the classical situation outlined at the beginning of this section does not lead to a qualitative change in the limit distributions, which is due to the rapid decay of the tails of geometric distributions. It does lead, however, to a change in the shift, which can be related to the fact that, with probability 1, the counting process $(N_n)_{n\in\mathbb{N}_0}$ has an asymptotically linear growth with slope $(b-1)/(2b)$.

# 4 Unsuccessful search in digital search trees

In our next example we again start with a sequence $(U_n)_{n\in\mathbb{N}}$ of independent random variables, now uniformly distributed on the unit interval. Let $T_n$ be the digital search tree associated with $U_1,\dots,U_n$; the algorithm is discussed in detail in (Mah92), Chapter 6, and (SF96), Chapter 7. We are interested in $I_n$, the depth of $U_n$ in $T_n$, which is the number of comparison needed for the last insertion into the tree.

For our approach it is essential that $I_n =_{\text{distr}} X_n$ for all $n\in\mathbb{N}$, where $X_n$ denotes the depth of the external node along a fixed path through the tree (note that the equality in distribution refers to the individual random variables, not to the whole sequences $(I_n)_{n\in\mathbb{N}}$, $(X_n)_{n\in\mathbb{N}}$). The stochastic process $(X_n)_{n\in\mathbb{N}_0}$ is a Markov chain on $\mathbb{N}_0$ with transition probabilities

$$p_{i,j} \;=\; \begin{cases} 2^{-i}, & \text{if } j=i+1,\\ 1-2^{-i}, & \text{if } j=i,\\ 0, & \text{otherwise,}\end{cases} \tag{7}$$

and start at $X_0=0$, the depth of the root node. Such a simple birth chain can alternatively be characterized by the fact that the holding times $T_0,T_1,\dots$ in the successively visited states $0,1,2,\dots$ are independent random variables, with $\mathcal{L}(T_j) = \mathrm{Geo}(2^{-j})$. Also, pure birth processes are counting processes, which provides a connection to renewal theory. However, in contrast to the standard renewal theoretic model the lifetimes are not identically distributed; instead, we have an exponential increase in distribution in the sense that $2^{-j}T_j$ converges in distribution as $j\to\infty$, the limit distribution being $\mathrm{Exp}(1)$. In (DG07) it is shown that this implies

$$\mathcal{L}\big(X_{n_k} - \lfloor \log_2 n_k\rfloor\big) \;\to\; Q_\eta \tag{8}$$

if $(n_k)_{k\in\mathbb{N}}$ is such that $n_k\to\infty$ and $\{\log_2 n_k\}\to\eta$ as $k\to\infty$. Here $Q_\eta$, $0\le\eta\le 1$, is the distribution of $\lfloor -\log_2 S+\eta\rfloor$, $S := \sum_{k=0}^{\infty} 2^{-k}Y_{\infty,k}$ and $Y_{\infty,k}$, $k\in\mathbb{N}_0$, are independent and identically distributed with $\mathcal{L}(Y_{\infty,1}) = \mathrm{Exp}(2)$. Again, the family of limit distributions arises from a fixed distribution $Q :=$

$\mathcal{L}(-\log_2 S)$ by shifts and subsequent discretization. As pointed out in (DG07), this can be used to obtain information about the tail behaviour of the limit distributions; for example,

$$Q_\eta([x,\infty)) \;=\; o\big(\exp(-\rho x^2)\big) \quad \text{as } x\to\infty, \quad \text{for all } \rho < (\log 2)/2. \tag{9}$$

Such representations can also be used to investigate continuity properties of the function $\eta \mapsto Q_\eta$. For example, if $Z$ has a density bounded by some finite constant $c$, then it is straightforward to prove that this function is then Lipschitz with respect to the Kolmogorov-Smirnov distance, i.e.

$$\sup_{j\in\mathbb{Z}} \big|Q_\eta((-\infty,j]) - Q_\zeta((-\infty,j])\big| \;\le\; c\,|\eta-\zeta| \quad \text{for all } \eta,\zeta\in[0,1].$$

For the density of interest in the present application some standard calculations lead to the upper bound $c = \exp(2)$. We mention in passing that a detailed knowledge of the density can also be used to obtain lower bounds for the total variation distance of the $Q_\eta$'s, for example.

It is tempting (especially in view of the approach to von Neumann addition outlined in the last section) to think of the periodicities as a result of the discretization of an underlying sequence $Y_1, Y_2, \dots$ of independent random variables, where $\mathcal{L}(Y_i) = \mathrm{Exp}(\lambda_i)$ with $\lambda_i := c(2^{-i})$, so that $T_i = \lceil Y_i \rceil$ for all $i\in\mathbb{N}$. With $Y_i$ instead of $T_i$ we obtain a continuous time Markov chain $\tilde{X} = (\tilde{X}_t)_{t\ge 0}$ of pure birth type with birth rates $q_{i,i+1} = \lambda_i$. However, it follows from the results in (DG07) that for any sequence $(n_k)_{k\in\mathbb{N}}$ with $\{\log_2 n_k\} \to \eta$ we obtain the same limit distribution for $\big(X_{n_k} - \lfloor \log_2 n_k \rfloor\big)$ and for $\big(\tilde{X}_{n_k} - \lfloor \log_2 n_k \rfloor\big)$ as $k\to\infty$. Hence, while we still have a family of limit distributions that can be obtained from a single distribution by shifts and discretization, the underlying mechanism here is, in contrast to the situation in the previous section, that the transition to discrete distributions comes from the sampling of a very slowly growing counting process.

The above model also appears in the context of approximate counting where the above interpretation as a pure birth process arises naturally; see (Fla85). The periodicities in the asymptotic distributional behaviour of unsuccessful search in digital search trees were discovered by (Lou87).

# 5 Loser election: The last few rounds

Suppose that a chairperson (loser) has to be chosen among $n$ candidates. These simultaneously throw coins in successive rounds and leave the competition if they obtain 'head'. We assume that the coin tosses are independent and that 'head' appears with some fixed probability $p$, $0 < p < 1$, which we regard as a parameter of the algorithm. If there is more than one candidate left in one of these rounds and if these all throw 'head' (a 'tie'), then the selection algorithm either ends up with more than one loser, or some modification is needed. This algorithm has attracted a lot of attention; see e.g. (LP06) and the references given there. A modification where additional rounds are introduced in the case of a final tie has been considered by (FMS96).

In probabilistic terms, the number of losers is the multiplicity of the maximum,

$$W_n \;=\; \#\{1\le j\le n : Y_j = M_n\}$$

where, as in Section 3, $M_n = \max\{Y_1,\dots,Y_n\}$ is the maximum of $n$ independent random variables $Y_1,\dots,Y_n$ with $\mathcal{L}(Y_j) = \mathrm{Geo}(p)$, $j = 1,\dots,n$. We define a family $\big\{Q_{p,\eta} : 0\le\eta<1\big\}$ of distributions by

$$Q_{p,\eta}\big(\{l\}\big) \;:=\; \frac{p^l}{l!}\sum_{j\in\mathbb{Z}} q^{l(j+\eta)}e^{-q^{j+\eta}} \quad \text{for all } l\in\mathbb{N}\,. \tag{10}$$

Then, Theorem 2 in (BG03) implies that

$$\lim_{n\to\infty} d_{\text{TV}}\big(\mathcal{L}(W_n), Q_{p,\eta_n}\big) \;=\; 0$$

with $\eta_n := \{\log_q n\}$; (Grü05) contains a simple proof. This situation seems to be noteworthy as it does not fit into the general framework that we considered in Section 2. To see this we argue as follows, for $p$ fixed: As the $Q_{p,\eta}$'s are all concentrated on $\mathbb{N}$ a relation of the type $Q_{p,\eta} = \mathcal{L}(\lceil Z+\eta\rceil)$ would mean that $P(Z\le 0)=0$. But then

$$Q_{p,\eta}(\{1\}) \;=\; P(\lceil Z+\eta\rceil = 1) \;=\; P(Z\le 1-\eta) \;\to\; 0 \quad \text{as } \eta\uparrow 1.$$

However, (10) implies

$$Q_{p,\eta}(\{1\}) \;\ge\; pq^{\eta}e^{-q^{\eta}} \;\ge\; pqe^{-1},$$

so we have a strictly positive lower bound that does not depend on $\eta$.

While (10) gives an explicit formula for the limit distributions it may be non-trivial to deduce even simple qualitative consequences from it. For example, it is 'probabilistically obvious' that the probability for a tie vanishes asymptotically as the parameter $p$ tends to 0. However, a proof of

$$\Big(\lim_{p\downarrow 0} Q_{p,\eta}(\{1\}) \;=\Big) \quad \lim_{p\downarrow 0} p\sum_{j\in\mathbb{Z}} q^{j+\eta}e^{-q^{j+\eta}} \;=\; 1 \tag{11}$$

seems to require some work. We now give a probabilistic construction that, as we hope, might augment our understanding of the algorithm and its distributional asymptotics, and which can also be used in connection with results such as (11).

As in Section 3 we could start with a sequence $(Z_n)_{n\in\mathbb{N}}$ of independent random variables with distribution $\text{Exp}(1)$ and use the fact that $(\lceil c(p)^{-1}Z_n\rceil)_{n\in\mathbb{N}}$ with $c(p)$ as in (6) is a sequence of independent random variables with distribution $\text{Geo}(p)$. We then have

$$W_n \;=_{\text{distr}}\; \#\big\{1\le j\le n:\ \lceil c(p)^{-1}Z_j\rceil = \lceil c(p)^{-1}\tilde{M}_n\rceil\big\}, \tag{12}$$

where again $\tilde{M}_n = \max\{Z_1,\dots,Z_n\}$. With this construction the sequence of maxima will be increasing, and subtraction of a suitable sequence of constants will give convergence in distribution (along subsequences, if we discretize), but this shifted sequence will not converge almost surely: The jumps of $(\tilde{M}_n)_{n\in\mathbb{N}}$ are more and more spaced out, but they are independent and exponentially distributed with mean 1. Now we note that for the multiplicity (and also for the value) of the maximum we do not need the full sample $Z_1,\dots,Z_n$, the (decreasing) order statistics $Z_{(n:1)},\dots,Z_{(n:n)}$ are enough: $W_n\ge j$ is equivalent to $\lceil c(p)^{-1}Z_{(n:j)}\rceil = \lceil c(p)^{-1}M_n\rceil$. For the transition

$$\big(Z_{(n:1)},\dots,Z_{(n:n)}\big) \;\mapsto\; \big(Z_{(n+1:1)},\dots,Z_{(n+1:n)},Z_{(n+1:n+1)}\big) \tag{13}$$

there is a simple distributional representation, based on a sequence $(V_n)_{n\in\mathbb{N}}$ of independent random variables, where now $\mathcal{L}(V_n) = \text{Exp}(n)$: We start with $Z_{(1:1)} = V_1$ and, for $n>1$, we put

$$Z_{(n+1:i)} = Z_{(n:i)} + V_{n+1} \;\text{ for } i=1,\dots,n, \quad Z_{(n+1:n+1)} = V_{n+1}. \tag{14}$$

This is the Sukhatme-Rényi representation of exponential order statistics, e.g. (SW86), p.721. In this representation we have $M_n = \sum_{i=1}^n V_i$, and the centred sequence $(M_n - H_n)_{n\in\mathbb{N}}$ is martingale that is easily seen to be $L^2$-bounded and hence converges almost surely and in (quadratic) mean to a random variable $M_\infty$. Of course, $M_\infty + \gamma$, with $\gamma := \lim_{n\to\infty}(H_n - \log n)$ denoting Euler's constant, has the standard Gumbel distribution mentioned in Section 3.

We now consider the counting process $N = (N_t)_{t\ge 0}$,

$$N_t := 1 + \inf\Big\{ n \in \mathbb{N} : \sum_{j=1}^n V_j \ge t \Big\} \quad \text{for all } t \ge 0, \tag{15}$$

associated with the sequence $(V_n)_{n\in\mathbb{N}}$. We write $N(t)$ instead of $N_t$ if this is typographically more convenient, and we extend the counting process to the negative halfline by setting $N_t = 0$ for $t < 0$. The process $N$ is again a Markov chain of pure birth type, but in contrast to Section 4 the holding times (or life times in a renewal theoretic interpretation) are now stochastically decreasing rather than increasing.

Within this framework we obtain a construction that reflects the behaviour of the algorithm over the last few rounds. To be specific, let $T_{n,0}$ be the number of candidates left when the game is over; in particular, $T_{n,0} = 1$ if and only if $W_n = 1$. For $j \in \mathbb{N}$ let $T_{n,j}$ be the number of candidates at the beginning of the last but $j$th round if we start with $n$ players. For example, with $n = 10$ players and with $2, 2, 0, 3, 3$ heads in the successive rounds we have $(T_{10,0}, T_{10,1}, T_{10,2}) = (0, 3, 6)$, whereas $2, 4, 0, 3, 1$ leads to $(T_{10,0}, T_{10,1}, T_{10,2}) = (1, 4, 4)$; we have $W_{10} = 3$ in the first and $W_{10} = 1$ in the second situation.

The following result shows that, again along suitable subsequences $(n_k)_{k\in\mathbb{N}}$, we have convergence of the $T_n$-processes in the sense of convergence in distribution of the finite segments $(T_{n,0}, T_{n,1}, \dots, T_{n,l})$ for every fixed $l \in \mathbb{N}_0$.

**Theorem 2** *Let $N = (N_t)_{t\ge 0}$ be the counting process associated with a sequence $(V_j)_{j\in\mathbb{N}}$ of independent random variables, with $V_j \sim \mathrm{Exp}(j)$ for all $j \in \mathbb{N}$. Then*

$$Z := \lim_{n\to\infty} \Big( \sum_{j=2}^n V_j - \log(n) \Big) \tag{16}$$

*exists almost surely. Further, let $(n_k)_{k\in\mathbb{N}} \subset \mathbb{N}$ be such that $n_k \to \infty$ and $\{c(p)^{-1}\log(n_k)\} \to \eta$ as $k \to \infty$. Then, for any fixed $l \in \mathbb{N}_0$,*

$$(T_{n_k,0}, T_{n_k,1}, \dots, T_{n_k,l}) \to_{\mathit{distr}} \big( N(\tau(p,\eta)), N(\tau(p,\eta) + c(p)), \dots, N(\tau(p,\eta) + l \cdot c(p)) \big), \tag{17}$$

*with*

$$\tau(p,\eta) := V_1 + c(p)\Big( \Big\{ \frac{Z}{c(p)} + \eta \Big\} - 1 \Big). \tag{18}$$

*In particular, for all $\eta \in [0,1]$ and all $p \in (0,1)$,*

$$Q_{p,\eta} = \mathcal{L}(W_{\infty,p,\eta}), \;\; \textit{with} \;\; W_{\infty,p,\eta} := \begin{cases} 1, & \textit{if } N(\tau(p,\eta)) = 1, \\ N(\tau(p,\eta) + c(p)), & \textit{otherwise}. \end{cases} \tag{19}$$

*Proof.* The above martingale argument also applies if we start with $j = 2$ instead of $j = 1$, so (16) holds.

With $S_n := V_1 + \cdots + V_n$ we have that $N(S_n - kc(p))$ is the number of participants after $k$ rounds, $k = 1, 2, \ldots$, and the number of rounds required is

$$L_n \;:=\; \min\bigl\{k \in \mathbb{N} :\; N(S_n - kc(p)) \in \{0,1\}\bigr\}.$$

This gives the following distributional representation of the number of participants after the last few rounds,

$$(T_{n,0}, T_{n,1}, \ldots, T_{n,l}) \;=_{\text{distr}}\; \bigl(N(S_n - L_n c(p)), N(S_n - (L_n - 1)c(p)), \ldots, N(S_n - (L_n - l)c(p))\bigr).$$

Hence it is enough to show that along subsequences $(n_k)_{k\in\mathbb{N}}$ with $\{c(p)^{-1}\log(n_k)\} \to \eta$

$$S_{n_k} - L_{n_k} c(p) \;\to\; V_1 \;+\; c(p)\Bigl(\Bigl\{\frac{Z}{c(p)} + \eta\Bigr\} - 1\Bigr) \tag{20}$$

almost surely as $k \to \infty$. Now it follows from the construction of $N$ that

$$L_n \;=\; \left\lceil \frac{V_2 + \cdots + V_n}{c(p)} \right\rceil \;=\; \frac{V_2 + \cdots + V_n}{c(p)} \;-\; \left\{ \frac{V_2 + \cdots + V_n}{c(p)} \right\} + 1$$

with probability 1, and $V_2 + \cdots + V_n - \log n \to Z$ together with the property of the subsequence yield

$$\frac{V_2 + \cdots + V_{n_k}}{c(p)} \;-\; \left\lfloor \frac{\log n_k}{c(p)} \right\rfloor \;\to\; \frac{Z}{c(p)} + \eta.$$

Taken together these imply (20), which completes the proof. □

Again, the various limit distributions (for fixed $p$) arise from a single distribution by operations that incorporate a shift and rounding at an early stage (as in Section 3), but we now also sample a counting process (as in Section 4). Moreover, (17), (18) and (19) respectively display the distributions for varying success probabilities $p$ as functions of one fixed random object $N$. The representation could therefore also be used to investigate the behaviour of the various limit distributions (for fixed $\eta$) as $p$ tends to 0 or 1. For example, we can now answer the question raised in connection with (11), and in fact obtain the exact rate of convergence to 0 for the probability of a tie as the parameter $p$ of the algorithm tends to 0.

**Corollary 1**

$$\lim_{p\downarrow 0} \frac{1}{p} Q_{p,\eta}(\{1\}^{\mathrm{c}}) \;=\; \frac{1}{2} \quad \textit{for all } \; \eta \in [0,1]. \tag{21}$$

*Proof.* For notational convenience we write $c = c(p) = -\log(1-p)$; clearly, $\lim_{p\to 0} p^{-1}c = 1$. We have $W_{\infty,p,\eta} \neq 1$ if and only if $\tau(p,\eta) < 0$. Hence, using the fact that $V_1$ and $Z$ are independent,

$$Q_{p,\eta}(\{1\}^{\mathrm{c}}) \;=\; P\bigl(V_1 < c(1 - \{c^{-1}Z + \eta\})\bigr) \;=\; \int (1 - e^{-cu})\, \mu_{c,\eta}(du),$$

where $\mu_{c,\eta}$ denotes the distribution of $1 - \{c^{-1}Z + \eta\}$. As $Z$ has a density the fractional parts $\{c^{-1}Z\}$ converge in distribution to the uniform distribution on the unit interval as $c \to 0$. A straightforward computation shows that that this then also holds for $1 - \{c^{-1}Z + \eta\}$. In particular,

$$\lim_{c\to 0} \int f(u)\,\mu_{c,\eta}(du) \;=\; \int_0^1 f(u)\,du$$

for all (bounded and) continuous functions $f : [0,1] \to \mathbb{R}$. This can be generalized to a family $(f_c)_{c\ge 0}$ instead of a fixed $f$ if the family is uniformly equicontinuous on the unit interval. With

$$f_c(u) = c^{-1}(1 - e^{-cu}) \text{ for } c > 0, \quad f_0(u) = u,$$

this property holds, so that

$$\lim_{p\to 0} \frac{1}{p} Q_{p,\eta}(\{1\}^c) \;=\; \lim_{c\to 0} \int \frac{1 - e^{-cu}}{c}\,\mu_{c,\eta}(du) \;=\; \int_0^1 u\,du \;=\; \frac{1}{2}\,.$$

□

The proof of the corollary can be extended to obtain asymptotic expansions for $Q_{p,\eta}(\{1\}^c)$ as $p \downarrow 0$. Note that the fluctuation disappears if we consider the behaviour of the limiting probability for a unique loser for small values of the basic parameter $p$.

The construction in this section can also be used to represent the number of rounds and, in fact, the joint distribution of the multiplicity and the number of rounds, which could serve as a starting point for the investigation of modifications of the algorithm that ensure that a single loser will be determined.

# 6 Remarks

## 6.1 Other representations

We have only treated cases where it is enough to shift the integer-valued random variables $X_n$ of interest by integer numbers $a_n$ in order to obtain non-trivial limit distributions along subsequences for $\mathcal{L}(X_n - a_n)$ as $n \to \infty$. This means that some measure of variability of $\mathcal{L}(X_n)$, such as the variance, or the mean absolute deviation, or the interquantile distances, is a bounded function of $n$. If this is not the case then scaling by another sequence $(b_n)_{n\in\mathbb{N}}$ with $b_n \to \infty$ may be necessary, so that instead of $X_n - a_n$ we now consider $\tilde{X}_n := (X_n - a_n)/b_n$ (it is then irrelevant whether $a_n$ is an integer or not). Again, we may have that $\mathcal{L}(\tilde{X}_n)$ does not converge, but that there is convergence along subsequences. Many such cases are known to arise in the analysis of algorithms, and often the family of limit distributions $Q_\eta$, $\eta \in M$, can be generated from a single complex-valued random variable $Z$ via

$$Q_\eta \;=\; \mathcal{L}\big(\mathrm{Re}(e^{i\eta}Z)\big) \quad \text{for all } \eta \in M. \tag{22}$$

Examples of this situation can be found in connection with $m$-ary search trees, see (CP04), in connection with urn models, see (Jan04a), and in connection with fragmentation trees, see (JN06). Hence, in contrast to the examples considered in the previous sections, where we shift and discretize, we now rotate and project.

### 6.2 Periodicity for partial sums

Theorems of the type

$$\lim_{n\to\infty} d\big(\mathcal{L}(\tilde{X}_n), Q_{\eta_n}\big) \;=\; 0 \tag{23}$$

with a family $\{Q_\eta : \eta \in M\}$ of limit points also appear, often under the label 'merge theorem', in the classical probability setup where sums $X_n = \sum_{j=1}^n Y_j$ of independent and identically distributed random variables $Y_j$, $j \in \mathbb{N}$, are considered. A particularly interesting case is S. Csörgö's analysis of the St. Petersburg paradox, where

$$P(Y_j = 2^l) \;=\; 2^{-l-1} \quad \text{for all } l \in \mathbb{N}_0, \tag{24}$$

see (CM02). In this situation it turns out that $\mathcal{L}(\tilde{X}_n)$ with $\tilde{X}_n := n^{-1}S_n - \log n$ is tight, with a family of limit points $Q_\eta$, $\eta \in [0,1]$, and $\eta_n = \{\log_2 n\}$. It may be interesting to investigate the distributional asymptotics of the renewal process $N = (N_t)_{t\ge 0}$, $N_t = \sup\{n \in \mathbb{N}_0 : X_n \le t\}$, with $X_0 := 0$, for lifetime distributions such as (24). This has a connection to fragmentation trees, where the behaviour along a path can be related to renewal processes by simply taking logarithms. For $m$-ary search trees it is known that the limiting periodicities vanish if the depth of the tree along a fixed path is considered, see (Jan04b), because the logarithm of a random variable with a beta distribution has finite mean—which, of course, is not the case in (24).

## References

[BG03] F. Thomas Bruss and Rudolf Grübel. On the multiplicity of the maximum in a discrete random sample. *Ann. Appl. Probab.*, 13:1252–1263, 2003.

[Bil68] Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons Inc., New York, 1968.

[Bil86] Patrick Billingsley. *Probability and Measure*. John Wiley & Sons Inc., New York, second edition, 1986.

[BvN46] H.H. Burks, A.W. Goldstine and J. von Neumann. *Preliminary discussion of the logical design of an electronic computing instrument*. Inst. for Advanced Study Report, Princeton, 1946.

[CM02] S. Csörgö and Z. Megyesi. Merging to semistable laws. *Teor. Veroyatnost. i Primenen.*, 47(1):90–109, 2002.

[CP04] Brigitte Chauvin and Nicolas Pouyanne. $m$-ary search trees when $m \ge 27$: a strong asymptotics for the space requirements. *Random Structures Algorithms*, 24(2):133–154, 2004.

[DG07] Florian Dennert and Rudolf Grübel. Renewals for exponentially increasing lifetimes, with an application to digital search trees. *Ann. Appl. Probab.*, 17:676–687, 2007.

[Fla85] Philippe Flajolet. Approximate counting: a detailed analysis. *BIT*, 25:113–134, 1985.

[FMS96] James Allen Fill, Hosam M. Mahmoud, and Wojciech Szpankowski. On the distribution for the duration of a randomized leader election algorithm. *Ann. Appl. Probab.*, 6:1260–1283, 1996.

[GR01] Rudolf Grübel and Anke Reimers. On the number of iterations required by von Neumann addition. *Theor. Inform. Appl.*, 35(2):187–206, 2001.

[Grü05] Rudolf Grübel. A hooray for Poisson approximation. In *2005 International Conference on Analysis of Algorithms*, Discrete Math. Theor. Comput. Sci. Proc., AD, pages 181–191 (electronic). Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2005.

[Jan04a] Svante Janson. Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Process. Appl.*, 110(2):177–245, 2004.

[Jan04b] Svante Janson. One-sided interval trees. *J. Iranian Statistical Soc.*, 3:149–164, 2004.

[Jan06] Svante Janson. Rounding of continuous random variables and oscillatory asymptotics. *Ann. Probab.*, 34:1807–1826, 2006.

[JN06] Svante Janson and Ralph Neininger. The size of random fragmentation trees. Preprint, 2006.

[Lou87] G. Louchard. Exact and asymptotic distributions in digital and binary search trees. *RAIRO Inform. Théor. Appl.*, 21:479–495, 1987.

[LP06] Guy Louchard and Helmut Prodinger. The asymmetric leader election algorithm: Another approach. Preprint, 2006.

[Mah92] Hosam M. Mahmoud. *Evolution of Random Search Trees*. John Wiley & Sons Inc., New York, 1992.

[Pro04] Helmut Prodinger. Periodic oscillations in the analysis of algorithms and their cancellations. *J. Iranian Statistical Soc.*, 3:251–270, 2004.

[Sco85] N.R. Scott. *Computer Number Systems & Arithmetic*. Prentice-Hall, New Jersey, 1985.

[SF96] Robert Sedgewick and Philippe Flajolet. *An Introduction to the Analysis of Algorithms*. Addison-Wesley, Reading, 1996.

[SW86] Galen R. Shorack and Jon A. Wellner. *Empirical Processes with Applications to Statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986.