Hoare's selection algorithm:
a Markov chain approach

Rudolf Grübel
Universität Hannover

We obtain bounds for the distribution of the number of comparisons needed by Hoare's randomized selection algorithm FIND and give a new proof for Grübel and Rösler's (1996) result on the convergence of this distribution. Our approch is based on the construction and analysis of a suitable associated Markov chain. Some numerical results for the quantiles of the limit distributions are included, leading for example to the statement that, for a set S with n elements and n large, FIND will need with probability 0.9 about $4.72 \cdot n$ comparisons to find the median of S.

AMS 1991 subject classifications. Primary 60J20; secondary 68P10.
Keywords and phrases. Randomized algorithms, stochastic ordering, convergence in distribution, Markov chains.

1. Introduction. Hoare's (1961) selection algorithm FIND finds the l^{th} smallest of a set $S \subset \mathbb{R}$ with n elements, $1 \le l \le n$. The variant to be discussed here proceeds as follows: If $n = 1$ the algorithm returns the only element of S. If $n > 1$, an element x is chosen uniformly at random from S, and the sets $S_- = \{y \in S : y < x\}$ and $S_+ = \{y \in S : y > x\}$ are determined, requiring $n-1$ comparisons. Let $m = |S_-|$ be the size of the set of smaller elements. If $m \ge l$ then continue with FIND(S_-, l). If $m = l - 1$ then return x. If $m < l - 1$ then continue with FIND$(S_+, l - m - 1)$. Obviously, the set S is reduced by at least one element in each recursion step and the algorithm thus terminates after at most $n - 1$ recursions.

Interest is in the total number $C_{n,l}$ of comparisons needed as this number essentially determines the time required by the algorithm. Devroye (1984) gave upper bounds for $P(C_{n,l} > nz)$ which hold uniformly in n and l and decrease exponentially in z. In Grübel and Rösler (1996), to which paper we also refer for related material and a more detailed discussion of the background, it was proved that $C_{n,l_n}/n$ converges in distribution as $n \to \infty$ if $\lim_{n\to\infty} l_n/n = t \in [0,1]$.

We will give a considerably simpler proof of this result. In Grübel and Rösler (1996) the random variables $C_{n,l}$, $1 \le l \le n$, for a slightly different version of the algorithm were considered simultaneously, and it was shown that the stochastic processes $(C_{n,\lceil nt \rceil}/n)_{0<t\le 1}$ converge in distribution as $n \to \infty$. Our new proof starts with the observation that the pairs (S, l) arising in the successive recursion steps constitute a Markov chain. For the counting of comparisons we may reduce

the state space of this chain to the pair (i,l), where i is the size of S; this does not destroy the Markov property. Indeed, if (i,l) is the current state, then the next state will be selected uniformly at random from the set

$$\big\{(1,1)\big\} \cup \big\{(i-k,l-k) : k=1,\dots,l-1\big\} \cup \big\{(i-k,l) : k=1,\dots,i-l\big\},$$

irrespective of the previous history of the process. If $n, i_1, i_2, \dots, 1, 1, 1 \dots$ is the sequence of first components of the successive states of the chain then $C_{n,l}$ is the sum of the values $n-1, i_1-1, i_2-1, \dots, 0, 0, 0 \dots$. The basic idea of the proof is that, after proper normalization, the sequence of chains arising as $n \to \infty$ converges to a continuous state space Markov chain; see Figure 1 below. To show that this convergence entails the convergence of $C_{n,l_n}/n$ we use a suitable almost sure construction that allows for a pathwise analysis.

This approach also leads to new results. One of these closes a 'conceptual gap' in Grübel and Rösler (1996): it was shown there that the tails of the limit distribution decrease at a faster than exponential rate. This does not permit a direct comparison with Devroye's (1984) exponential bounds, however, as the latter refer to the distribution of $C_{n,l}/n$ for finite n. Theorem 1 below shows that the stochastic upper bound derived in Grübel and Rösler (1996) for the limit distribution in fact applies to the distribution of $C_{n,l}/n$ too, for all $n \in \mathbb{N}$ and all $l \in \{1,\dots,n\}$. This seems to be of some interest, firstly because a recent result of Goldie and Grübel (1996) can be applied to obtain the exact logarithmic rate of decrease of the tail probabilities of the upper bound (which is $-z \log z$), and secondly because the bound is tight in a sense which will be made precise below. Also, via the stochastic upper bound, we obtain a simple and explicit, non-asymptotic numerical upper bound for $P(C_{nl} > nz)$ in Theorem 2.

Theorem 3 is the convergence result mentioned above; a stochastic lower bound for the limit distributions (which depend on the limit of l_n/n) is given in Theorem 4. In Theorem 5 these limit distributions are related to each other.

We write $\mathcal{L}(X)$ for the distribution ('law') of a random quantity X and $\mathcal{L}(X|Y)$ for the conditional distribution of X given Y. The symbol '$\to_{\text{distr}}$' denotes convergence in distribution. Also, $X \le_{\text{distr}} Y$ for nonnegative random variables X, Y indicates that X is smaller than or equal to Y in stochastic order, i.e. $P(X > z) \le P(Y > z)$ for all $z \ge 0$. With both '$\to_{\text{distr}}$' and '$\le_{\text{distr}}$' we will occasionally follow the common sloppy practice of mixing random variables and distributions when notationally convenient. We write $\star$ for convolution, δ_c for the unit mass in $c \in \mathbb{R}_+$ and 1_A for the indicator function of the set A. Finally, $X \sim \text{unif}(a,b)$ abbreviates the statement that X is a real-valued random variable which is uniformly distributed on the interval (a,b).

2. Main results. A discrete time Markov chain can be regarded as a collection of stochastic processes indexed by an initial state. The distributional aspects of this collection are completely specified by the state space I and the transition kernel P of the Markov chain.

Let $I := \{(x,y) \in \mathbb{R}_+^2 : 0 \le y \le x\}$ and for each $(x,y) \in I$ let $P\big((x,y),\,\cdot\,\big)$ be the distribution of ξ,

$$\xi \;:=\; \begin{cases} (x - Ux, y - Ux), & \text{if } \; Ux \le y, \\ (Ux, y), & \text{if } \; Ux > y, \end{cases} \qquad \text{with } U \sim \text{unif}(0,1). \tag{1}$$

This defines a transition kernel P on I; let $Z = (Z_m)_{m \in \mathbb{N}_0}$ with $Z_m = (X_m, Y_m)$ be a Markov chain with state space I and transition kernel P.

For each $n \in \mathbb{N}$ let

$$I_n \;:=\; \Bigl\{ \Bigl(\frac{i}{n}, \frac{l}{n}\Bigr) : i, l \in \mathbb{N},\; l \le i \Bigr\},$$

and for each $(i/n, l/n) \in I_n$ let $P_n\bigl((i/n, l/n), \cdot\bigr)$ be the discrete uniform distribution on

$$\Bigl\{ \Bigl(\frac{1}{n}, \frac{1}{n}\Bigr) \Bigr\} \cup \Bigl\{ \Bigl(\frac{i-k}{n}, \frac{l-k}{n}\Bigr) : k = 1, \ldots, l-1 \Bigr\} \cup \Bigl\{ \Bigl(\frac{i-k}{n}, \frac{l}{n}\Bigr) : k = 1, \ldots, i-l \Bigr\}.$$

This defines a transition kernel on I_n. Let $Z^{(n)} = (Z_m^{(n)})_{m \in \mathbb{N}_0}$ with $Z_m^{(n)} = (X_m^{(n)}, Y_m^{(n)})$ be a Markov chain with state space I_n and transition kernel P_n. Figure 1 illustrates the basic transition mechanism for the n-chain and its continuous counterpart.

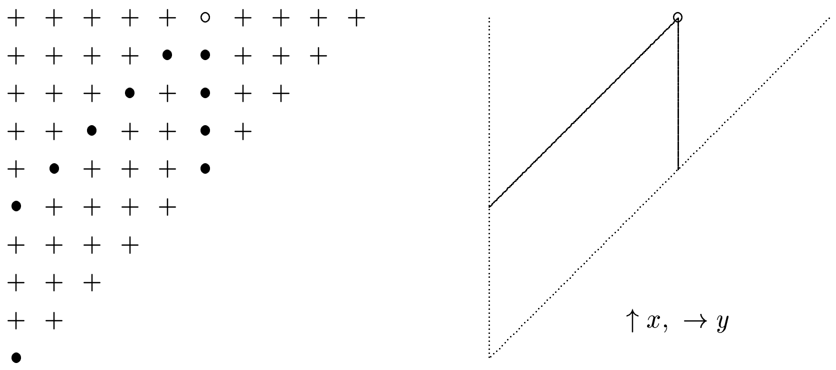

FIGURE 1 *Transitions of $Z^{(n)}$ and Z*
○: current state; •, —: potential next states

The following relation provides the connection to the algorithm FIND, it displays $C_{n,l}$ as a function of $Z^{(n)}$:

$$\mathcal{L}\Bigl(\frac{1}{n} C_{n,l}\Bigr) \;=\; \mathcal{L}\Bigl(\sum_{m=0}^{\infty} \Bigl(X_m^{(n)} - \frac{1}{n} \Bigr) \Bigm| Z_0^{(n)} = \Bigl(1, \frac{l}{n}\Bigr) \Bigr). \tag{2}$$

This, together with a suitable construction of the $Z^{(n)}$-chains, leads to the following uniform stochastic upper bound for $C_{n,l}$. Proofs are collected in the next section.

THEOREM 1 *Let* $(V_m)_{m\in\mathbb{N}}$ *be a sequence of independent random variables with distribution* unif$(1/2,1)$*, and let* $W := 1+\sum_{m=1}^{\infty}\prod_{r=1}^{m} V_r$*. Then*

$$\frac{1}{n}\,C_{n,l} \;\le_{\mathrm{distr}}\; W \quad \textit{for all } n\in\mathbb{N},\ l\in\{1,\dots,n\}.$$

The statement of the theorem can also be written as

$$\sup_{1\le l\le n} P(C_{nl}\ge z) \;\le\; P(W\ge z) \quad \text{for all } \ n\in\mathbb{N},\ z\ge 0.$$

A similar statement with 'the supremum inside P', i.e. on the tails of $\sup_{1\le l\le n} C_{nl}$, would involve the joint distribution of $C_{n1},\dots,C_{nn}$. A brief discussion of this problem is given in Section 4.3.

Sums over cumulative products of i.i.d. sequences, such as W in Theorem 1, have been studied by various authors, see e.g. Goldie and Grübel (1996) and the references given there. They appear in insurance mathematics where the above W arises as the 'perpetuity' associated with unif$(1/2,1)$. Goldie and Grübel (1996) have recently investigated the tail behaviour of such perpetuities and obtained the following result:

$$\lim_{z\to\infty}\frac{1}{z\log z}\,\log P(W\ge z) \;=\; -1. \tag{3}$$

This together with Theorem 1 implies that $\sup_{n\in\mathbb{N},1\le l\le n} P(C_{n,l} > nz)$ decreases with increasing z at a faster than exponential rate. Of course, for any finite n, the support of the distribution of $C_{n,l}/n$ has a finite upper endpoint; this bound, however, increases with n. A closer analysis of W results in the following non-asymptotic bound.

THEOREM 2 *For all* $z\ge 0$*,* $n\in\mathbb{N}$ *and* $1\le l\le n$*,*

$$P(C_{nl}\ge nz) \;\le\; \exp(z+z\log 4 - z\log z).$$

The same construction as for the proof of Theorem 1, now carried out simultaneously for all $Z^{(n)}$-chains and the process Z, shows that $Z^{(n)}$ converges to Z in a sense sufficiently strong to allow for application of the functional appearing in (2).

THEOREM 3 *Let* $(l_n)_{n\in\mathbb{N}}$ *be a sequence of integers with* $1\le l_n\le n$ *such that* l_n/n *converges to a limit* $t\in[0,1]$ *as* $n\to\infty$*. Then, as* $n\to\infty$*,*

$$\frac{1}{n}\,C_{n,l_n} \;\to_{\mathrm{distr}}\; \mathcal{L}\Big(\sum_{m=0}^{\infty} X_m \,\Big|\, Z_0=(1,t)\Big).$$

Let $Q_{x,y} := \mathcal{L}\big(\sum_{m=0}^{\infty} X_m \,|\, Z_0=(x,y)\big)$. For us the distributions $Q_{1,t}$, $0\le t\le 1$, are of particular importance as they arise as the limits in Theorem 3. The convergence in distribution implies $Q_{1,t}\le_{\mathrm{distr}} W$ with W as in Theorem 1, see also Theorem 9 in Grübel and Rösler (1996). The next theorem gives a related stochastic lower bound.

THEOREM 4 *Let* $W' := 1 + \sum_{m=1}^{\infty} \prod_{r=1}^{m} U_r$ *where* $(U_r)_{r\in\mathbb{N}}$ *is a sequence of independent random variables with distribution* $\mathrm{unif}(0,1)$*. Then* $Q_{1,t} \geq_{\mathrm{distr}} W'$ *for all* $t \in [0,1]$*.*

From the definition of Z it follows easily that $Q_{1,0} = Q_{1,1} = \mathcal{L}(W')$, hence the lower bound in Theorem 4 is tight. This lower bound is the perpetuity associated with $\mathrm{unif}(0,1)$. The result of Goldie and Grübel (1996) can be applied to W' too, and leads to the same tail behaviour, i.e. (3) also holds for W'. Hence the stochastic interval obtained for the distributions $Q_{1,t}$ is small in the sense that the tail behaviour of upper and lower bound coincide if measured on a logarithmic scale; see also Section 4.2 below.

The Markov property of $(Z_m)_{m\in\mathbb{N}_0}$ together with a scaling property of the associated transition kernel can be used to obtain an integral equation that relates the distributions $Q_{1,t}$, $0 \leq t \leq 1$, to each other. This in turn can be used to obtain information on the limit distributions, numerically or otherwise. In order to make this precise we need the functions $T_c : \mathbb{R}_+ \to \mathbb{R}_+$, $c > 0$, defined by $T_c(x) := cx$. Further let $T_c(Q)$ denote the image of the distribution Q on $\mathbb{R}_+$ under T_c (if X is a random variable with distribution Q then $T_c(Q)$ is the distribution of cX).

THEOREM 5 *For all* $t \in [0,1]$*,*

$$Q_{1,t} \;=\; \int_0^t \delta_1 \star T_{1-u}\big(Q_{1,\frac{t-u}{1-u}}\big)\, du \;+\; \int_t^1 \delta_1 \star T_u\big(Q_{1,\frac{t}{u}}\big)\, du.$$

The formula in Theorem 5 can be used to obtain integral equations for the moments of the limit distributions, e.g. for $m_1(t) := \int x\, Q_{1,t}(dx)$ we get

$$m_1(t) \;=\; 1 \;+\; \int_0^t (1-u)\, m_1\Big(\frac{t-u}{1-u}\Big)\, du \;+\; \int_t^1 u\, m_1\Big(\frac{t}{u}\Big)\, du$$

which is solved by $m_1(t) = 2 - 2t\log t - 2(1-t)\log(1-t)$ (see also Grübel and Rösler (1996), Theorem 11). Paulsen (1995) recently solved the analogous integral equations for higher moments of $Q_{1,t}$ and in particular obtained an explicit expression for the variance. In view of the bound in Theorem 1 the convergence in Theorem 3 implies the existence and convergence of all moments, i.e.

$$\lim_{n\to\infty} E\Big(\frac{1}{n^k}\, C_{n,l_n}^k\Big) \;=\; \int x^k\, Q_{1,t}(dx) \;<\; \infty$$

for all $k \in \mathbb{N}$ if $l_n/n \to t$. In Section 4 we will use Theorem 5 to derive an integral equation for the distribution functions of $Q_{1,t}$, $0 \leq t \leq 1$, and obtain some numerical results for the associated quantiles.

3. Proofs. For Theorem 1 and Theorem 3 we use a suitable almost sure construction.

Let $(U_m)_{m\in\mathbb{N}}$ be a sequence of independent, unif$(0,1)$-distributed random variables. We define $(Z_m)_{m\in\mathbb{N}_0}$ with $Z_m = (X_m, Y_m) \in I$ recursively by $Z_0 := (1,t)$ and

$$Z_{m+1} \;:=\; \begin{cases} (X_m - U_{m+1}X_m, Y_m - U_{m+1}X_m), & \text{if } U_{m+1}X_m \le Y_m, \\ (U_{m+1}X_m, Y_m), & \text{if } U_{m+1}X_m > Y_m. \end{cases} \tag{4}$$

Obviously, $(Z_m)_{m\in\mathbb{N}_0}$ is a Markov chain with transition kernel P and start at $(1,t)$, i.e. we have used the sequence $(U_m)_{m\in\mathbb{N}}$ to construct a suitable version of the prospective limit process.

We now use the same sequence to construct suitable processes $(Z_m^{(n)})_{m\in\mathbb{N}_0}$, $n \in \mathbb{N}$. To this end we need the auxiliary functions

$$h_n : \Big\{\frac{k}{n} : 1 \le k \le n\Big\} \times [0,1) \to \Big\{\frac{k}{n} : 1 \le k \le n\Big\}, \quad h_n\Big(\frac{i}{n}, u\Big) \;:=\; \frac{1}{n}\lceil iu \rceil.$$

Obviously, if $U \sim$ unif$(0,1)$ then $h_n(i/n, U)$ has the discrete uniform distribution on $\{k/n : 1 \le k \le i\}$. We now define $(Z_m^{(n)})_{m\in\mathbb{N}_0}$, $Z_m^{(n)} = (X_m^{(n)}, Y_m^{(n)}) \in I_n$ recursively by $Z_0^{(n)} := (1, l_n/n)$ and

$$Z_{m+1}^{(n)} \;:=\; \begin{cases} (X_m^{(n)} - \zeta, Y_m^{(n)} - \zeta), & \text{if } \zeta < Y_m^{(n)}, \\ (1/n, 1/n), & \text{if } \zeta = Y_m^{(n)}, \\ (\zeta - 1/n, Y_m^{(n)}), & \text{if } \zeta > Y_m^{(n)}, \end{cases} \qquad \text{with } \zeta := h_n(X_m^{(n)}, U_{m+1}).$$

It is easy to see that $(Z_m^{(n)})_{m\in\mathbb{N}_0}$ is a Markov chain with transition kernel P_n and start at $(1, l_n/n)$. In view of (2) the statement of Theorem 3 will follow if we can show that

$$\sum_{m=0}^{\infty} \Big(X_m^{(n)} - \frac{1}{n}\Big) \;\to\; \sum_{m=0}^{\infty} X_m \qquad \text{almost surely as } n \to \infty. \tag{5}$$

The main tool for the proof of this will be a uniform upper bound on the ratios of successive X-values. Let $V_m := \max\{U_m, 1 - U_m\}$ for all $m \in \mathbb{N}$; note that $(V_m)_{m\in\mathbb{N}}$ is an i.i.d. sequence of unif$(1/2,1)$-distributed random variables. It is immediate from the construction of $(Z_m)_{m\in\mathbb{N}_0}$ that

$$\frac{X_{m+1}}{X_m} \;\le\; V_{m+1} \qquad \text{for all } m \in \mathbb{N}_0. \tag{6}$$

For the discretized versions we use the elementary inequalities

$$\frac{i - 1 - \lceil iu \rceil}{i-1} \;\le\; 1 - u, \quad \frac{\lceil iu \rceil - 2}{i-1} \;\le\; u \quad \text{for all } 0 < u < 1,\; i = 2, 3, \ldots, n$$

to obtain

$$\frac{X_{m+1}^{(n)} - \frac{1}{n}}{X_m^{(n)} - \frac{1}{n}} \;\le\; V_{m+1} \quad \text{for all } m \in \mathbb{N}_0,\; n \in \mathbb{N} \;\text{ such that }\; X_m^{(n)} > \frac{1}{n}. \tag{7}$$

From this last inequality the statement of Theorem 1 follows immediately.

What happens as $n \to \infty$? We have

$$X_{m+1}^{(n)} \;=\; \phi_n\big(X_m^{(n)}, Y_m^{(n)}, U_{m+1}\big), \quad X_{m+1} \;=\; \phi\big(X_m, Y_m, U_{m+1}\big),$$

with

$$\begin{aligned} \phi_n(x,y,u) \;&=\; \Big(x - \frac{1}{n}\lceil nux\rceil\Big)\, 1_{(0,y)}\Big(\frac{1}{n}\lceil nux\rceil\Big) \\ &\qquad\quad +\; \frac{1}{n}\big(\lceil nux\rceil - 1\big)\, 1_{(y,1)}\Big(\frac{1}{n}\lceil nux\rceil\Big) \;+\; \frac{1}{n}\, 1_{\{y\}}\Big(\frac{1}{n}\lceil nux\rceil\Big), \\ \phi(x,y,u) \;&=\; (x-ux)\, 1_{(0,y]}(ux) \;+\; ux\, 1_{(y,1)}(ux). \end{aligned}$$

Obviously, if $(x_n, y_n) \to (x,y) \in I$ with $ux \neq y$, then $\phi_n(x_n,y_n,u) \to \phi(x,y,u)$. Together with a similar analysis for the second component this shows that the implication

$$(X_m^{(n)}, Y_m^{(n)}) \to (X_m, Y_m) \text{ a.s.} \quad\Longrightarrow\quad (X_{m+1}^{(n)}, Y_{m+1}^{(n)}) \to (X_{m+1}, Y_{m+1}) \text{ a.s.}$$

holds for all $m \in \mathbb{N}_0$. Hence, as $(X_0^{(n)}, Y_0^{(n)}) \to (X_0, Y_0)$ by construction,

$$X_m^{(n)} - \frac{1}{n} \;\to\; X_m \qquad \text{almost surely as } n \to \infty \text{ for all } m \in \mathbb{N}. \tag{8}$$

It is an elementary matter to show that W is finite with probability 1 (in fact, we have already mentioned that the tails of W decrease at a faster than exponential rate). If W is finite, then the bounds (6) and (7) for the ratios of the X-values show that we may apply dominated convergence to conclude that convergence in (8) holds even after summation over m. Hence (5) holds and Theorem 3 is proved.

For the proof of Theorem 4 we cannot use a pathwise comparison as in the above proof for Theorem 1 and Theorem 3, we have to compare distributions. Note that $X \le_{\text{distr}} Y$ is equivalent to $Eh(X) \le Eh(Y)$ for all non-decreasing functions $h : \mathbb{R}_+ \to [0,1]$. Using (4) is is straightforward to check that

$$\frac{X_{m+1}}{X_m} \;\ge\; (1-\xi_m)\, U_{m+1} \;+\; \xi_m\, (1 - U_{m+1}) \quad \text{with} \quad \xi_m := 1_{(0.5,1)}\Big(\frac{Y_m}{X_m}\Big),$$

which in turn implies

$$\mathcal{L}\big(X_{m+1} \,\big|\, Z_0, \dots, Z_m\big) \;\ge_{\text{distr}}\; \text{unif}(0, X_m) \;\text{ for all } m \in \mathbb{N}_0, \tag{9}$$

with $Z_m = (X_m, Y_m)$. Now let $(U'_m)_{m\in\mathbb{N}}$ be an i.i.d. sequence of unif$(0,1)$-variables and define $(X'_m)_{m\in\mathbb{N}}$ by $X'_0 := X_0$, $X'_{m+1} := U'_{m+1} X'_m$ for all $m \in \mathbb{N}$. We claim that

$$\sum_{m=0}^{M} \alpha_m X_m \;\ge_{\text{distr}}\; \sum_{m=0}^{M} \alpha_m X'_m \quad \text{for all } \alpha_1, \dots, \alpha_M \ge 0 \tag{10}$$

holds for all $M \in \mathbb{N}_0$. This is obviously true for $M = 0$. Assume now that (10) is true for some $M \in \mathbb{N}_0$ and let $h : \mathbb{R}_+ \to [0,1]$ be non-decreasing. Then, for arbitrary $\alpha_1, \dots, \alpha_{M+1} \ge 0$,

$$\begin{aligned}
&E\,h\Bigl(\sum_{m=0}^{M+1} \alpha_m X_m\Bigr) \\
&\quad = \int\!\!\int h\Bigl(\sum_{m=0}^{M} \alpha_m X_m + \alpha_{M+1} X_{M+1}\Bigr)\, d\mathcal{L}(X_{M+1} \,|\, Z_0, \dots, Z_M)\, d\mathcal{L}(Z_0, \dots, Z_M) \\
&\quad \ge \int\!\!\int_0^1 h\Bigl(\sum_{m=0}^{M} \alpha_m X_m + u\alpha_{M+1} X_M\Bigr)\, du\; d\mathcal{L}(Z_0, \dots, Z_M) \\
&\quad = \int_0^1 E\,h\Bigl(\sum_{m=0}^{M-1} \alpha_m X_m + (\alpha_M + u\alpha_{M+1}) X_M\Bigr)\, du \\
&\quad \ge \int_0^1 E\,h\Bigl(\sum_{m=0}^{M-1} \alpha_m X'_m + (\alpha_M + u\alpha_{M+1}) X'_M\Bigr)\, du \;=\; E\,h\Bigl(\sum_{m=0}^{M+1} \alpha_m X'_m\Bigr),
\end{aligned}$$

where we have used (9) in the first and (10) in the second inequality. Hence (10) holds for all $M \in \mathbb{N}_0$. Using (10) with all α's equal to 1 we obtain for all $z \ge 0$

$$\begin{aligned}
P\Bigl(\sum_{m=0}^{\infty} X_m > z\Bigr) \;&=\; \lim_{M\to\infty} P\Bigl(\sum_{m=0}^{M} X_m > z\Bigr) \\
&\ge\; \lim_{M\to\infty} P\Bigl(\sum_{m=0}^{M} X'_m > z\Bigr) \;=\; P\Bigl(\sum_{m=0}^{\infty} X'_m > z\Bigr),
\end{aligned}$$

which completes the proof of Theorem 4.

To prove Theorem 5 we decompose with respect to the value of Z_1 and obtain

$$\begin{aligned}
Q_{x,y} \;&=\; \int \mathcal{L}\Bigl(x + \sum_{m=1}^{\infty} X_m \,\Big|\, Z_1\Bigr)\, d\mathcal{L}\bigl(Z_1 \,|\, Z_0 = (x,y)\bigr) \\
&=\; \int_0^{y/x} \delta_x \star Q_{x-ux,\,y-ux}\, du \;+\; \int_{y/x}^{1} \delta_x \star Q_{ux,y}\, du.
\end{aligned}$$

From the definition of the transition kernel P (or, more explicitly, on using the almost sure construction introduced at the beginning of this section) it follows that

$$Q_{cx,cy} \;=\; T_c\bigl(Q_{x,y}\bigr) \qquad \text{for all } (x,y) \in I,\ c > 0.$$

Combining these two statements we obtain the assertion of Theorem 5.

It remains to prove Theorem 2. The argument given for the asympotic upper bound in Goldie and Grübel (1996, p.473) shows that the moment generating function $M(\theta) = E\exp(\theta W)$ of W satisfies $M(\theta) \le \exp\bigl(C\exp(\theta)\bigr)$ for all $\theta \ge 0$ if C is such that

$$2\int_{1/2}^{1} \exp\bigl(C\exp(\theta u)\bigr)\, du \;\le\; \exp\bigl(C\exp(\theta) - \theta\bigr) \quad \text{for all } \theta \ge 0. \tag{11}$$

A standard argument based on Markov's inequality yields the assertion of Theorem 2 if we can show that (11) holds with $C = 4$. To achieve this, let $f_l(\theta)$ and $f_r(\theta)$ denote the left and right hand side respectively of (11) with $C = 4$. Because of $f_l(0) = f_r(0)$ it is enough to show that

$$\frac{d}{d\theta}\big(\theta f_l(\theta)\big) \;\le\; \frac{d}{d\theta}\big(\theta f_r(\theta)\big) \qquad \text{for all } \theta > 0. \tag{12}$$

Using the substitution $x := \exp\big(4\exp(\theta u)\big)$ we obtain

$$\theta f_l(\theta) \;=\; 2\int_{\exp(4\exp(\theta/2))}^{\exp(4\exp(\theta))} \frac{1}{\log(x)}\,dx$$

so that

$$\frac{d}{d\theta}\big(\theta f_l(\theta)\big) \;=\; 2\exp\big(4\exp(\theta)\big) \;-\; \exp\big(4\exp(\theta/2)\big).$$

Letting $y := \exp(\theta/2)$, multiplying by $y^2\exp(-4y^2)$ and collecting terms we see that (12) follows if

$$g(y) \;:=\; -2y^2 \;+\; y^2\exp\big(4y(1-y)\big) \;+\; 1 \;+\; 8y^2\log(y) \;-\; 2\log(y) \;\ge\; 0$$

for all $y > 1$. It is easy to check that $g(1) = g'(1) = 0$, so this would in turn follow from

$$g''(y) \;\ge\; 0 \qquad \text{for all } y > 1. \tag{13}$$

We have

$$g''(y) \;=\; 20 \;+\; p(y)\,\exp\big(4y(1-y)\big) \;+\; 16\,\log(y) \;+\; \frac{2}{y^2}$$

with $p(y) \;:=\; 64y^4 - 64y^3 - 24y^2 + 16y + 2$. Elementary arguments show that $p(y) \ge p(1) = -6$ for all $y \ge 1$, hence (13) follows on using that $\exp\big(4y(1-y)\big) \le 1$ on the range of interest.

We note in passing that these methods can also be used to show that $C = 4$ is the best possible constant in (11).

4. Miscellaneous complements.

4.1 Let $F(t,x) = Q_{1,t}([0,x])$ be the distribution function associated with $Q_{1,t}$. Theorem 4 leads to the integral equation

$$F(t,x) \;=\; \int_0^t F\Big(\frac{t-u}{1-u}, \frac{(x-1)^+}{1-u}\Big)\,du \;+\; \int_t^1 F\Big(\frac{t}{u}, \frac{(x-1)^+}{u}\Big)\,du \tag{14}$$

where $x^+ := \max\{x, 0\}$. There seems to be little hope to solve this explicitly, but (14) can be put to numerical use: discretization together with the trapezoidal rule leads to an approximate version of (14) which in turn can be solved to any desired degree of precision by iteration. Table 1 presents the values obtained in this manner for the 0.5-, 0.9- and 0.99-quantiles of $Q_{1,t}$ for some t-values (note that

t	0.50	0.90	0.99
0.00	1.89	2.97	4.11
0.10	2.52	3.82	5.15
0.25	3.01	4.41	5.81
0.50	3.27	4.72	6.16

TABLE 1 *Quantiles of* $Q_{1,t}$

$Q_{1,t} = Q_{1,1-t}$). Section 2 contains an explicit expression for the expectation $m_1(t)$ associated with $Q_{1,t}$, hence upper bounds for these quantiles can be obtained on using Markov's inequality: $Q_{1,t}\big([\alpha m_1(t), \infty)\big) \le 1/\alpha$. With $t = 0.5$ and $1/\alpha = 0.01$ this would lead to the bound 338.63, which overestimates the value 6.16 from Table 1 by a considerable amount; Theorem 2 gives the bound 14.84.

4.2 The upper and lower bounds on $Q_{1,t}$ from Section 2 are close to each other in the tails. The numerical procedure outlined in the preceding subsection can also be used in connection with perpetuities; Figure 2 shows the distribution functions associated with the bounds and with $Q_{1,0.5}$. Note that the bound in Theorem 2 is of interest only from $z = 4e \approx 10.873$ onwards.

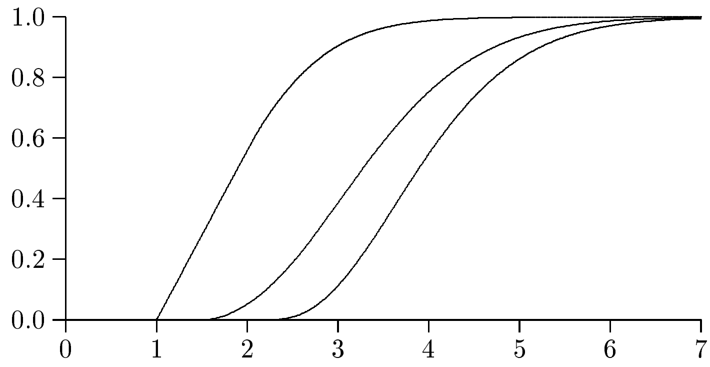

FIGURE 2 $Q_{1,0.5}$: *distribution function and bounds*

4.3 The results of Grübel and Rösler (1996) can be used to obtain the joint limit distribution of $C_{n,l_n}/n$ and $C_{n,k_n}/n$ if $l_n/n \to s$ and $k_n/n \to t$, for arbitrary $s, t \in [0,1]$. Such a result needs an assumption on how the selection procedures depend on each other once the sets obtained in the course of the l_n-search and the k_n-search split; in Grübel and Rösler (1996) these selections were assumed to be independent. Alternatively, we could base all selections in the m^{th} recursion level on the same U_m, $U_m \sim \text{unif}(0,1)$, by selecting s_{mi} with $i := \lceil n_m U_m \rceil$, if $\{s_{m1}, \dots, s_{mn_m}\}$ with $s_{m1} < \dots < s_{mn_m}$ is the set under consideration at the entry into this level (dependent selection). In practice FIND will operate on arrays (vectors) rather than sets. As long as we consider one l only, or in the case of independent selection, the difference between arrays and sets is irrelevant due to the permutation invariance of discrete uniform distributions; with dependent selection, however, the joint distribution of the number of comparisons needed to find the smallest and the largest item in S, for example, will depend on the order of the entries of S. Interestingly, this dependence on the input disappears if we consider the maximum of all $C_l(S)$, with l varying from 1 to the length of S. It is this quantity that is of interest in connection with a worst case analysis, and with dependent selection we obtain

$$\lim_{n\to\infty} \sup_{1\le l\le n} \frac{1}{n} C_{n,l} \;=\; 1 \;+\; \sum_{m=1}^{\infty} \prod_{k=1}^{m} \max\{U_k, 1-U_k\}. \tag{15}$$

Loosely speaking, the worst case arises if our opponent, knowing how we will make our choices, hides the element we wish to find in such a way that it is in the larger one of the sets S_-, S_+ in each recursion step. We have argued before that the bound in Theorem 1 is tight if interest is in the tail behaviour of the distributions, (15) supplements this: the right hand side is identical in distribution to the stochastic upper bound in Theorem 1.

References

Devroye, L. (1984) Exponential bounds for the running time of a selection algorithm. *J. of Computer and System Sciences* **29**, 1-7.

Goldie, C.M. and Grübel, R. (1996) Perpetuities with thin tails. *Adv. in Applied Prob.***28**, 463-480.

Grübel, R. and Rösler, U. (1996) Asymptotic distribution theory for Hoare's selection algorithm. *Adv. in Applied Prob.* **28**, 252-269.

Hoare, C.A.R. (1961) Algorithm 65, FIND. *Communications of the ACM* **4**, 321-322.

Paulsen, V. (1995) The moments of FIND. Preprint, Universität Kiel.

INSTITUT FÜR MATHEMATISCHE STOCHASTIK
UNIVERSITÄT HANNOVER
POSTFACH 60 09
D-30060 HANNOVER

e-mail: rgrubel@stochastik.uni-hannover.de