Some remarks on random subsets of interval

Jacek Cichon, Marek Klonowski, Przemyslaw Kobylanski,
Łukasz Krzywiecki, Bartłomiej Rozanski,
Paweł Zielinski

2005
Some remarks on random subsets of interval

Jacek Cichoń, Marek Klonowski, Przemysław Kobylański, Łukasz Krzywiecki, Bartłomiej Różański, Paweł Zieliński

ABSTRACT. In this paper we compare two methods of generation of finite random subsets of the interval $[0, 1]$. Both methods are based on an independent family of random variables uniformly distributed on $[0, 1]$. The first method uses the values of random variables directly as the splitting point. The second one used $i$-th variable only for choosing an interval, which is then split precisely into two subintervals of equal lengths. We discuss some statistical properties of these methods.

The first method corresponds to the well known routing protocol Chord. The second, with better statistical properties, may be used for some modification of Chord and, in fact, corresponds to one-dimensional version of another well known P2P protocol CAN. A modification of Chord based on the idea is very gentle and fairly practical.

1. Introduction

If $X$ is a random variable then we denote by $E[X]$ its mean value and we denote by $\text{var}[X]$ its variance. If $A$ is a subset of a probability space $\Omega$ then we denote by $1_A$ its characteristic function, e.g. the function which takes value 1 on $A$ and 0 on the set $\Omega \setminus A$.

Let us consider a sequence $\zeta_1, \ldots, \zeta_n$ of independent uniformly distributed random variable with values in $[0, 1]$. Then the set $\{\zeta_1(\omega), \ldots, \zeta_n(\omega)\}$ is a random subset of the interval $[0, 1]$. This method of generation of random subset is called a uniform split (see [7]). We now describe a model. Let $\lambda^\alpha$ denote the $\alpha$-dimensional Lebesgue measure on the space $\mathbb{R}^\alpha$. For $n \in \mathbb{N}$ and $\alpha > 0$ we put

$$\Omega_{\alpha,n} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : (\forall i)(x_i \geq 0) \land \sum_{i=1}^{n} x_i \leq \alpha\}.$$

Then $\lambda^n(\Omega_{\alpha,n}) = \frac{\alpha^n}{n!}$. We consider the set $\Omega_n = \Omega_{1,n}$ as a probability space with the probability defined as

$$P_n(A) = n! \cdot \lambda^n(A).$$

If $x \in \Omega_n$ then we put $x_{n+1} = 1 - (x_1 + \ldots + x_n)$ and $r(x, k) = \sum_{i \leq k} x_i$. Notice that $0 \leq r(x, 1) \leq \ldots \leq r(x, n) \leq 1$. We identify an element $x \in \Omega_n$ with the random subset of $[0, 1]$

$$r(x) = \{r(x, k) : k = 1, \ldots, n\}$$

of $[0, 1]$ which contains the point 0. One can easily show that this model describes the uniform split.

2000 Mathematics Subject Classification. Primary 05A16; Secondary 05A19.

Key words and phrases. random set, finitary combinatorics, Chord, CAN, P2P, consistent hashing.
Let us consider another method of generation of random subsets of $[0, 1]$. We start from the set $\{0, 1\}$. Suppose that we have chosen points $\{P_1, \ldots, P_n\}$ and a new point is to be added to this set. We choose a random number $x \in [0, 1]$ and find two neighbors $P, Q$ of $x$ from points chosen so far and a new point $P_{n+1}$ is put precisely in the middle between $P$ and $Q$. We call this method a \textit{binary split}. Let us describe this process more formally. So, let $\Omega$ be a probability space and let us fix a sequence $(\xi_1, \ldots, \xi_n)$ of independent random variables on $\Omega$ uniformly distributed in the interval $[0, 1]$; each $\omega \in \Omega$ generates a random subset of $[0, 1]$ obtained from the sequence $(\xi_1(\omega), \ldots, \xi_n(\omega))$ for sequential splitting intervals into two subintervals of equal length. By $x_1(\omega), \ldots, x_{n+1}(\omega)$ we denote the lengths of intervals generated by sequence $(\xi_1(\omega), \ldots, \xi_n(\omega))$.

Let $\mathcal{P}$ be any random method of generating a random subset of interval $[0, 1]$ of cardinality $n$. The set $\mathcal{P}(\omega)$ defines a sequence $(\chi(\omega), \ldots, \chi_n(\omega))$ of lengths of consecutive intervals. Notice that $x_1(\omega) + \ldots + x_{n+1}(\omega) = 1$, hence $\frac{x_1(\omega) + \ldots + x_{n+1}(\omega)}{n+1} = \frac{1}{n+1}$. We define

$$\text{var}_n(\mathcal{P}) = E\left[\frac{\sum_{i=1}^{n+1}(x_i - \frac{1}{n+1})^2}{n+1}\right].$$

and $\text{std}_n(\mathcal{P}) = \sqrt{\text{var}_n(\mathcal{P})}$. We treat the number $\text{std}_n(\mathcal{P})$ as a measure of non-uniformity of distribution of points from a random set of cardinality $n$ generated by process $\mathcal{P}$. It is easy to check that

$$(1.1) \quad \text{var}_n(\mathcal{P}) = \frac{1}{n+1}(E\left[\sum_{i=1}^{n+1}x_i^2\right] - \frac{1}{n+1}).$$

Suppose that we have some subset $\{a_1, \ldots, a_n\}$ of $[0, 1]$, where $a_1 < a_2 < \ldots < a_n < a_{n+1} = 1$. Suppose that we are choosing a random point $\zeta$ from $[0, 1]$ according to the uniform distribution on $[0, 1]$ and we choose interval $[a_{i-1}, a_i]$ such that $\zeta \in [a_{i-1}, a_i]$. This interval we call a \textit{randomly uniformly chosen interval}.

\textbf{Theorem 1.1.} Let $\mathcal{P}_n$ be a random method of generation of a random subsets of interval $[0, 1]$. Then the number

$$\text{ELRI}_n(\mathcal{P}) = E\left[\sum_{i=1}^{n+1}x_i^2\right]$$

is the expected value of randomly uniformly chosen interval.

\textbf{Proof.} Let $\zeta$ be a random number of uniform distribution on $[0, 1]$. Then we have

$$\text{ELRI}_n(\mathcal{P}) = \sum_{i=1}^{n+1} E[x_i : \zeta \in I_i]P(\zeta \in I_i) = E\left[\sum_{i=1}^{n+1}x_i^2\right].$$

\hfill $\square$

\section{Uniform split}

In this section we discuss some basic properties of uniform split of the interval $[0, 1]$. This process we denote by \textit{unif}. First we consider the variance of distribution of lengths of intervals.

\textbf{Theorem 2.1.} $\text{ELRI}_n(\text{unif}) = \frac{2}{n+2}$
PROOF. The result follows from direct calculations:

\[
ELR_{n}(\text{unif}) = E\left[\sum_{i=1}^{n+1} x_{i}^{2}\right] = \int_{\Omega_{n}} (x_{1}^{2} + \ldots + x_{n+1}^{2})dP(\omega) = \\
(n+1) \int_{\Omega_{n}} x_{1}^{2}dP(\omega) = (n+1)\int_{\Omega_{n}} x^{2}n^{-1}n^{-1}(\Omega_{1-n^{-1}})dx = \\
(n+1)n! \int_{0}^{1} x^{2} n^{-1} (1-x)^{n-1} dx = n(n+1) 2\Gamma(n) \frac{2\Gamma(n)}{\Gamma(3+n)} = \frac{2}{n+2}.
\]

\[
\square
\]

COROLLARY 2.1. \(\text{var}_{n}(\text{unif}) = \frac{n}{(1+n)^{2}(2+n)}\)

Notice that the average of distances between two consecutive points in a random subset of \([0, 1]\) of cardinality \(n\) is \(\frac{1}{n+1}\) and that \(\text{std}_{n}(\text{unif}) \sim \frac{1}{n}\). Therefore we may expect that there are large oscillations of lengths of intervals in random sets generated by the uniform split.

In fact the uniform split has the following properties:

1. there exists precisely one interval of length less than \(\frac{1}{n}\) with high probability (w.h.p.)
2. there exists precisely one interval of length bigger than \(\frac{\ln n}{n}\) w.h.p.
3. there are approximately \(\sqrt{n}\) of intervals of lengths less than \(\frac{\sqrt{n}}{n}\) w.h.p.
4. there are approximately \(\sqrt{n}\) of intervals of lengths bigger than \(\frac{\ln n}{2n}\) w.h.p.

We omit standard proofs of these facts. Let us conclude that the lengths of intervals generated by the uniform split lay (w.h.p.) in the interval \([\frac{1}{n+1}, \frac{\ln n}{2n}]\).

2.1. Statistical estimator. Let us look at the value \(ELR_{n}(\mathcal{P})\) once again. We consider the probability space \(\Omega_{n} \times [0, 1]\) with the product measure \(P = P_{n} \otimes \lambda\) and let \(L_{n} : \Omega_{n} \times [0, 1] \rightarrow [0, 1]\) be the random variable defined by the formula

\[
L_{n}(x, \alpha) = \max\{k : r(x, k) \leq \alpha\} + 1.
\]

By direct use of the Fubini theorem we deduce that

\[
P(L_{n} < t) = 1 - (n+1)(1-t)^{n} - (1-t)^{n+1},
\]

so the density \(\varphi_{n}\) of the random variable \(L_{n}\) is

\[
\varphi_{L_{n}}(t) = (n+1)n(1-t)^{n-1}.
\]

From this we can easily calculate that \(E[L_{n}] = \frac{2}{n+2}\), from which we can once again get \(ELR_{n}(\text{unif}) = \frac{2}{n+2}\).

We now discuss a method of estimation the number of nodes in a uniform split. We assume that we choose a random sample modelled as a sequence \(x_{1}, \ldots, x_{k}\) of equally distributed, independent random variables with density \(\varphi_{L_{n}}\).

THEOREM 2.2. Suppose that \(x_{1}, \ldots, x_{k}\) is a random sample from a distribution with density \(\varphi_{L_{n}}\). Let

\[
\hat{n}(x_{1}, \ldots, x_{k}) = \frac{1}{2} + C + \sqrt{C^2 + 1},
\]

where \(C = \ln \frac{1}{n!1_{n\geq (1-\alpha)}}\). Then

1. \(\hat{n}\) is an unbiased estimator (i.e. \(E[\hat{n}] = n\)),
2. \(\hat{n}\) is asymptotically (with respect to \(k\)) normal,
(3) the variance of the asymptotic distribution is equal

\[ D^2[\hat{n}] = \frac{1}{k} \frac{n^2(n+1)^2}{n^2 + (n+1)^2}. \]

**Proof.** Let us consider the likelihood function (see [3])

\[ H(x_1, \ldots, x_n, n) = \ln \prod_{i=1}^{k} \varphi_{\hat{t}_n}(x_i). \]

The only positive solution of the equation \( \partial H / \partial n = 0 \) is \( n = -\frac{1}{2} + C + \sqrt{C^2 + 1} \), where \( C = (-k) / \ln \prod_{i=1}^{k} (1 - x_i) \).

Let \( f(x, n) = \ln \varphi_{\hat{t}_n}(x) \). Then

\[ \frac{\partial^2 f}{\partial n^2} = -(n^2 + (n+1)^2) / n^2(n+1)^2, \]

so

\[ D^2[\hat{n}] = \frac{1}{k} \int \frac{\partial^2 f}{\partial n^2} \varphi_{\hat{t}_n}(x) dx = \frac{1}{k} \frac{n^2(n+1)^2}{n^2 + (n+1)^2}. \]

Notice that \(-\frac{1}{2} + C + \sqrt{C^2 + 1} \approx 2C \) for large \( n \). Therefore for large \( n \) we may use the following approximation

\[ \hat{n} \approx \frac{-2k}{\ln \prod_{i=1}^{k} (1 - x_i)} \]

Moreover \( D^2[\hat{n}] \approx \frac{1}{2k}. \)

### 3. Binary split

We discuss in this section the binary split model of generation of random subsets of the interval \([0, 1]\). We denote this process by \( \text{bin} \). Our main goal is to calculate the number \( \text{var}_n(\text{bin}) \). We put \( f_0 = 1 \) and \( f_n = ELRI_n(\text{bin}) \) for \( n > 0 \).

**Lemma 3.1.** For all natural numbers \( n \) we have

\[ f_{n+1} = \frac{1}{2n+1} \sum_{k=0}^{n} \binom{n}{k} f_k. \]

**Proof.** Let us consider the sequence \( \{\xi_1, \ldots, \xi_{2n+1}\} \) and \( \omega \in \Omega \). At the beginning the number \( \xi_1(\omega) \) splits the interval \([0, 1]\) into two equal parts \([0, 0.5] \) and \([0.5, 1]\). Let \( A = \{ i > 1 : \xi_i(\omega) < 0.5 \} \) and \( B = \{ i > 1 : \xi_i(\omega) > 0.5 \} \). Then variables \( \{\xi_i : i \in A\} \) split only the interval \([0, 0.5] \) and variables \( \{\xi_i : i \in B\} \) split only the interval \([0.5, 1]\). Note that \( \{2\xi_i : i \in A\} \) split the interval \([0, 1]\). Hence \( \mathbf{E}(2\xi_i)_{i \in A} = f_{|A|} \). Similar observation holds for the the sequence \( (2\xi_i - 1)_{i \in B} \). Therefore we have

\[ f_{n+1} = \sum_{A \subseteq \{2, \ldots, n+1\}} \left( \frac{1}{2} f_{|A|} \right) \left( \frac{1}{2} \right)^{|A|} n^{-|A|} = \frac{1}{2n+2} \sum_{k=0}^{n} \binom{n}{k} (f_k + f_{n-k}) = \frac{1}{2n+2} \sum_{k=0}^{n} \binom{n}{k} f_k. \]

\[ \square \]
Before we formulate next result, we define numbers

\[ L_n = \prod_{j=n}^{\infty} \left( 1 - \left( \frac{1}{2} \right)^j \right). \]

Then \( L_1 \approx 0.2888 \) and the inequalities \( 1 - \frac{1}{2^n} < L_n < 1 - \frac{1}{2^{n+1}} \) holds for each \( n \geq 1 \).

**Lemma 3.2.** \( f_n = \sum_{m \geq 0} \left( \frac{1}{2} \right)^m (1 - \left( \frac{1}{2} \right)^m)^n L_{m+1} \).

**Proof.** Let us consider the generating function of the sequence \( (f_n)_{n \geq 0} \):

\[ x(t) = \sum_{n \geq 0} f_n \frac{t^n}{n!}. \]

It is not difficult to check that the sequence satisfies the following functional equation

\[ 2x'(2t) = x(t)e^t, \]

i.e. \( x'(t) = \frac{1}{2} x(t)e^t \). If we put \( X(t) = x(t)e^{-t} \) then we obtain a slightly simpler equation:

\[ X'(t) = \frac{1}{2} X \left( \frac{t}{2} \right) - X(t). \]

This equation can be solved explicitly, and we obtain

\[ X(t) = \sum_{n \geq 0} \frac{t^n}{n!} (-1)^n \prod_{k=1}^{n} \left( 1 - \left( \frac{1}{2} \right)^k \right). \]

Since \( x(t) = X(t)e^t \) we obtain

\[ f_n = \sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) (-1)^k \prod_{j=1}^{k} \left( 1 - \left( \frac{1}{2} \right)^j \right). \]

It is hard to calculate accurately this formula since it contains large coefficients with alternating signs. Therefore we transform this formula into more suitable form. After putting in the Euler formula (see [6])

\[ \prod_{k=a+1}^{\infty} \frac{1}{1 - q^k z} = \sum_{n \geq 0} \frac{z^n}{\prod_{k=1}^{a} (1 - q^k)} \]

\[ z = q^{a+1} \] and \( q = \frac{1}{2} \) we get

\[ \frac{1}{\prod_{k=a+1}^{\infty} (1 - \left( \frac{1}{2} \right)^k)} = \sum_{n \geq 0} \frac{\left( \frac{1}{2} \right)^{(a+1)n}}{\prod_{k=1}^{a} (1 - \left( \frac{1}{2} \right)^k)}. \]

After multiplying both sides of this equality by \( L_1 \) we get

\[ \prod_{j=1}^{a} \left( 1 - \left( \frac{1}{2} \right)^j \right) = \sum_{n \geq 0} \left( \frac{1}{2} \right)^{(a+1)n} L_{n+1}. \]

Therefore we have

\[ f_n = \sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) (-1)^k \sum_{m \geq 0} \left( \frac{1}{2} \right)^{(k+1)m} L_{m+1} = \]

\[ = \sum_{m \geq 0} L_{m+1} \left( \frac{1}{2} \right)^m \sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) (-1)^k \left( \left( \frac{1}{2} \right)^m \right)^k. \]
This fact follows quickly from monotonicity of the function so the lemma is proved.

Notice that from Lemma 3.2 it follows that \( f_{n+1} < f_n \) for each \( n \). Let us consider the function

\[
\varphi_n(x) = \left( \frac{1}{2} \right)^x \left( 1 - \left( \frac{1}{2} \right)^x \right)^n
\]
defined on the interval \([0, \infty)\). The function \( \varphi_n \) has a global maximum at point \( \log(n+1) \) and \( \varphi_n(\log(n+1)) = \frac{1}{ne} + o\left( \frac{1}{n} \right) \). Notice that \( \sum_{m \geq 0} (\frac{1}{2})^m (1 - (\frac{1}{2})^m)^n = \sum_{m \geq 0} \varphi_n(m) \).

Moreover, \( \int_0^\infty \varphi_n(x) \, dx = \frac{1}{\ln(2)+\ln(2)}, \) from which we deduce that \( f_n = O\left( \frac{1}{n} \right) \).

**Lemma 3.3.** \( f_n = \sum_{m \geq 0} \frac{1}{2m} (1 - \frac{1}{2^m})^n + o\left( \frac{1}{n} \right) \)

**Proof.** The proof is done by simple estimation. First we show that

\[
\sum_{m=0}^{\log \sqrt{n}} \frac{1}{2m} (1 - \frac{1}{2^m})^n = o\left( \frac{1}{n} \right);
\]

This fact follows quickly from monotonicity of the function \( \varphi_n \) on the interval \([0, \log(n+1)]\). Namely,

\[
\varphi_n(\log \sqrt{n}) \approx \frac{1}{\sqrt{ne\sqrt{n}}},
\]

so

\[
\sum_{m=0}^{\log \sqrt{n}} \frac{1}{2m} (1 - \frac{1}{2^m})^n \leq \frac{\log \sqrt{n}}{\sqrt{ne\sqrt{n}}} \leq \frac{1}{e\sqrt{n}}
\]

and \( \frac{1}{e\sqrt{n}} = o\left( \frac{1}{n} \right) \). Observe that if \( k > \log \sqrt{n} \), then \( L_k > 1 - \frac{4}{\sqrt{n}} \) so, we have

\[
(1 - \frac{4}{\sqrt{n}}) \sum_{m=0}^{\log \sqrt{n}} \frac{1}{2m} (1 - \frac{1}{2^m})^n \leq \sum_{m=\log \sqrt{n}} L_{m+1} \leq \sum_{m=\log \sqrt{n}} \frac{1}{2m} (1 - \frac{1}{2^m})^n L_{m+1}.
\]

Since \( f_n = O\left( \frac{1}{n} \right) \), we have

\[
| \sum_{m=\log \sqrt{n}} \frac{1}{2m} (1 - \frac{1}{2^m})^n - \sum_{m=\log \sqrt{n}} \frac{1}{2m} (1 - \frac{1}{2^m})^n L_{m+1} | \leq \frac{C}{n\sqrt{n}} = o\left( \frac{1}{n} \right),
\]

so the lemma is proved.

**Lemma 3.4.** \( \sum_{k=0}^{\infty} \frac{1}{2k} (1 - \frac{1}{2^k})^n = \sum_{k=0}^{n} \binom{n}{l} (-1)^l \frac{1}{1 - (\frac{1}{2})^{1+k}} \)

**Proof.** The proof follows from the following transformations:

\[
\sum_{k \geq 0} \frac{1}{2k} (1 - \frac{1}{2^k})^n = \sum_{k \geq 0} \frac{1}{2k} \sum_{l=0}^{n} \binom{n}{l} (-1)^l \frac{1}{2^l} = \\
\sum_{l=0}^{n} \binom{n}{l} (-1)^l \sum_{k \geq 0} \frac{1}{2^{kl+k}} = \sum_{l=0}^{n} \binom{n}{l} (-1)^l \frac{1}{1 - (\frac{1}{2})^{1+l}}.
\]
THEOREM 3.1. The sequence $f_n$ satisfies

$$f_n = \frac{1}{\ln 2} + \frac{\omega(\log_2 n)}{n} + o\left(\frac{1}{n}\right),$$

where $\omega$ is a periodic function with period 1 such that $|\omega(x)| < 10^{-3}$.

In the proof of this theorem we use a standard method for the treatment of sums of that type, attributed to S.O. Rice by D.E. Knuth (see [6]).

PROOF. Let us denote, for simplicity,

$$s_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k \frac{1}{1 - \left(\frac{1}{2}\right)^{1+k}}$$

and let us consider the complex function

$$f(z) = \frac{1}{1 - \left(\frac{1}{2}\right)^{1-z}}.$$  

Then $f$ is a meromorphic function with single pools at points

$$x_k = 1 + \frac{2\pi i k}{\ln(2)},$$

where $k \in \mathbb{Z}$ and $i$ is the imaginary unit and $s_n = \sum_{k=0}^{n} (-1)^k f(z)$. Let $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ be the Euler beta function and let us observe that the function $B(n+1,z)$ has (single) pools at points $0, -1, \ldots, -n$.

Moreover

$$Res(B(n+1,z); z = -k) = (-1)^k \binom{n}{k}.$$  

Notice that the function $f$ is holomorphic on the half-plane $\Re(z) < 0.5$. Therefore

$$s_n = \sum_{k=-n}^{0} Res(B(n+1,z)f(z); z = -k).$$

Let us consider a big rectangle $C_k$ with end-point $\pm k \pm (2k+1)\pi i / \ln 2$, where $k > n$ is a big positive number. It is easy to check that

$$\lim_{k \to \infty} \oint_{C_k} B(n+1,z)f(z)dz = 0.$$  

In fact, it is sufficient to observe that $f(1 + (2k+1)\pi i / \ln 2) = 1/2$. Therefore, using Cauchy theorem, we get

$$s_n = -\sum \{Res[B(n+1,z)f(z); x = x_k]: k \in \mathbb{Z}\}.$$  

Next, it is also easy to check that

$$Res[B(n+1,z)f(z); z = 1] = -\frac{1}{(n+1)\ln 2}.$$  

This first residue gives us the first part of the approximation of the number $s$. Generally, we have

$$Res[B(n+1,z)f(z); z = x_k] = -\frac{\Gamma(1+n)\Gamma(1 + \frac{2\pi k}{\ln 2})}{\Gamma(2+n + \frac{2\pi k}{\ln 2})\ln 2}.$$
It is easy to see, that in our estimation we may take only into account the residues at point \(x_{-1}, x_0, x_1\). Therefore

\[
s_n = \frac{1}{(n+1) \ln 2} - \text{Res}[B(n+1, z)f(z); z = x_{-1}] + \text{Res}[B(n+1, z)f(z); z = x_1] + o\left(\frac{1}{n}\right).
\]

We use now the following approximation formula

\[
n^{b-a} \frac{\Gamma(z+a)}{\Gamma(z+b)} = 1 + \frac{(a-b)(a+b-1)}{2z} + O\left(\frac{1}{z^2}\right),
\]

and get

\[
\text{Res}[B(n+1, z)f(z); z = x_k] = -\Gamma(1 + \zeta) \frac{1}{n^{1+\zeta} \ln 2} \left(1 + \frac{(1+\zeta)(2+\zeta)}{2} + O\left(\frac{1}{n^2}\right)\right),
\]

where \(\zeta = x_k - 1\) and finally, after some simplifications, we obtain

\[
s_n \approx \frac{1}{(n+1) \ln 2} + \frac{1}{n} \left(0.000228 \cos(2\pi \log n) - 0.00057 \sin(2\pi \log n)\right) + o\left(\frac{1}{n}\right).
\]

\[\square\]

A crucial role in the last proof was played by three main pools of the function \(f(z) = 1/(1 - (1/2)^{1-\zeta})\): \(1, 1 + 2\pi i/\ln 2\) and \(1 - 2\pi i/\ln 2\). The first one, located at point 1, is responsible for the main part of the sequence \(s_n\), i.e., for the component \(1/n \ln 2\). The next two pools are responsible for relatively small oscillations of the sequence \(s_n\). The size of these oscillations is relatively small because \(\Im(1 + 2\pi i/\ln 2) \approx 10\) and the function \(\Gamma\) decreases rapidly when the imaginary part of a number grows. From the last theorem and equality 1.1 we obtain

**Corollary 3.1.** 
\[\text{var}_n(\text{bin}) = \frac{\sqrt{\pi}}{n} + o\left(\frac{\log n}{n}\right) + o\left(\frac{1}{n}\right)\]

Notice that \(\frac{1}{\ln 2} - 1 \approx 0.4427\), therefore \(\text{std}_n(\text{bin}) \sim \frac{0.665}{n}\), hence \(\text{std}_n(\text{bin})\) is especially smaller than the average length of intervals in uniform split model, which is \(\frac{1}{\ln 2}\). Therefore the binary split process gives a more uniform distribution of random points in interval [0, 1] than the uniform split.

Let us consider the random variable \(L_n\) defined as the length of the first interval in the binary split model (of length \(n\)) and let \(C_n = \frac{1}{L_n}\). Then the values of \(C_n\) are positive integers and the distribution of \(C_n\) coincides with the distribution of values of the R. Morris probabilistic counter (see [10]). A detailed analysis of \(C_n\) was done by P. Flajolet (see [2]) and later by P. Kirchenhofer and H. Prodinger (see [5]). They showed that

\[
E[C_n] \sim \log_2 n + \frac{\gamma}{\log_2 n} - \alpha + \frac{1}{2} + o(\log_2 n)
\]

where \(\alpha = \sum_{\ell=1}^{\ln 2 - 1} \frac{1}{2^\ell - 1}\) and \(o(\cdot)\) is a periodic function with period 1, mean 0 and amplitude less than \(10^{-4}\).

The above analysis may be applied to an arbitrary node. It is shown in [7] that the distribution of lengths of uniformly chosen interval in the binary split model coincides with the distribution of length of the first node.
4. On routing protocols

All our mathematical investigations discussed so far were motivated by investigations in computer science - namely by behavior of various P2P lookup protocols. One of them is Chord (see [12] and [8]). It belongs to the most popular P2P protocols for internet applications. Compared to other known P2P routing protocols, such as CAN ([11]), Pastry ([1]) or Kademlia ([9]), three features make Chord stay out from the competition: its simplicity, provable correctness and provable performance.

Since its presentation Chord received living interest and has been subject of deep evaluation along with scientific research of protocol aspects like e.g. network dynamics. As a result several changes to the protocol were proposed (see [8]) which eliminates some basic protocol weaknesses.

Chord network consists of nodes aligned on a directed identifier circle. Nodes’ main functionality is to provide network lookups. In addition to lookups, each party in Chord protocol has to perform maintenance jobs. These additional procedures preserve fast and accurate lookups even in case of frequent topology changes caused by node joins and leaving the protocol.

Chord routing is based on mapping of nodes and resources (shares) on a directed identifier circle. Mapping is done according to some hash function with uniform distribution. Node (computer) is placed in the identifier circle according to the hash value of its unique label - for example IP number. Upon network join node takes over responsibility for all shares that belong to the network interval between it and its predecessor in Chord topology. In original Chord this interval is split straightforward according to the hash function value, i.e. joining node is inserted at every location of its hash function identifier. In other words, the process of inserting new node on the identifier circle is equivalent with the uniform split process analyzed in Section 2. Hence, with high probability there will be many non-equal intervals resulting. Moreover, subsequent divisions may lead to vast discrepancy between interval sizes leaving some nodes with network load much larger ($O(n \log(n))$) than others.

Our proposal of improving Chord is based on a simple idea of dividing intervals into two subintervals of equal size - hence we propose to use the binary split process instead of uniform one. The better performance of this modification is explained by Corollary 3.1. We call the modified protocol Chord$^W$. Note that this modification is very subtle - it does not destroy any of fundamental property of Chord (simplicity, correctness, performance) and is easy for implementation. Simulations shows that Chord$^W$ has better distribution of nodes than Chord also after deletion of nodes, i.e. that the standard deviation of lengths of intervals after successive adding and deleting of nodes is smaller that $\frac{1}{n}$. Our experiment also shows that the statistical estimator of numbers of nodes from Section 2 behaves very well also for Chord$^W$.

Nodes in CAN protocol (see [11]) are placed in the hypercube $[0, 1)^d$ for some fixed $d \geq 1$. Each node $P$ takes a control over some hyper-perpendicular of the form $H^P = \prod_{i=1}^d (\alpha_i^P, \beta_i^P)$, where $0 \leq \alpha_i < \beta_i \leq 1$. When a new node $Q$ is to be added, a sequence $L(Q) = (h_1(Id_Q), \ldots, h_d(Id_Q))$ of values of independent hash function is calculated ($Id_Q$ is, for example, the IP of the node $Q$) and the unique node $P$ such that $L(Q) \in H^P$ is found; then the hyper-perpendicular $H^P$ is splitted into two hyper-perpendiculars $H_1$ and $H_2$ of the same volume. The node $P$ will be responsible for $H_1$ and the node $Q$ for $H_2$. Note that the distribution of volumes in CAN is precisely described by the binary split process. Hence the distribution of volumes in CAN does not depend on the dimension $d$ - of course, the dimension $d$ has influence on other properties of CAN, such as number of neighbors or the
length of search path. Therefore, we may treat $Chord^W$ as a one-dimensional version of CAN with modified look-up technique.

References


INSTITUTE OF MATHEMATICS AND COMPUTER SCIENCE, WROCLAW UNIVERSITY OF TECHNOLOGY, WYBRZEZE WSPÓŁNAŃSKIEGO 27, 50-370 WROCLAW, POLAND