Kybernetika

Ivan Kramosil

Independent and identically distributed pseudo-random samples

Kybernetika, Vol. 23 (1987), No. 1, 1--12

Persistent URL: `http://dml.cz/dmlcz/124771`

KYBERNETIKA — VOLUME 23 (1987), NUMBER 1

INDEPENDENT AND IDENTICALLY DISTRIBUTED PSEUDO-RANDOM SAMPLES

IVAN KRAMOSIL

The most simple algorithm used in order to transform a statistically independent identically distributed random sequence with equiprobable distribution into sequence with some given non-equiprobable distribution is proved to be applicable also in the case of pseudo-random sequences of high algorithmic complexity. The result enables to present a definition of i.i.d. pseudo-random sequences based immediately on the notion of algorithmic complexity.

1. INTRODUCTION

Consider an infinite, statistically independent and identically distributed (i.i.d.) sequence $X = X_1, X_2, \ldots$ of random variables, defined on a probability space $\langle \Omega, \mathcal{S}, P \rangle$ and taking their values in a finite set $\mathcal{A} = \{a_1, a_2, \ldots, a_r\}$. Set, for $i = 1, 2, \ldots, r$,

$$p_i = P(\{\omega : \omega \in \Omega, X_1(\omega) = a_i\}), \tag{1}$$

then $p = \langle p_1, p_2, \ldots, p_r \rangle$ is the probability distribution of X (corresponding to X). Realizations of i.i.d. random sequences with a given probability distribution play an important role in some stochastic computational decision making or simulation methods.

Let $\langle p_1, p_2, \ldots, p_r \rangle$ be a probability distribution over $\mathcal{A}$ and let $Y = Y_1, Y_2, \ldots$ be another i.i.d. sequence of random variables taking $\langle \Omega, \mathcal{S}, P \rangle$ into $\mathcal{B} = \{b_1, b_2, \ldots \ldots, b_m\}$ such that $P(\{\omega : \omega \in \Omega, Y_1(\omega) = b_j\}) = m^{-1}$ for each $j \leqq m$. If there are $n_1, n_2, \ldots, n_r \in \mathcal{N} = \{0, 1, 2, \ldots\}$ such that $\sum_{i=1}^{r} n_i = m$ and $p_i = n_i/m$, $i = 1, 2, \ldots, r$, then a realization $Y_1(\omega), Y_2(\omega), \ldots$ of Y can be transformed into a realization of i.i.d. X with probability distribution $\langle p_1, p_2, \ldots, p_r \rangle$, setting simply $X_i\,(\omega) = a_j$ iff $Y_i(\omega) = b_k$ and $\sum_{s=0}^{j-1} n_s < k \leqq \sum_{s=0}^{j} n_s$, $n_0 = 0$.

The infinite sequences of elements from $\mathscr{B}$, the initial segments of which are combinatorially complicated enough lest to be generated by programs substantially shorter than their lengths, are known to be good simulations of realizations of i.i.d. random sequences with equiprobable distribution. The aim of this paper is to investigate whether, in which sense and in which measure, the sequences obtained from such pseudo-random sequences over $\mathscr{B}$ by the factorization transformation can simulate i.i.d. sequences over $\mathscr{A}$ with probability distribution $\langle p_1, p_2, \ldots, p_r \rangle$.

2. PSEUDO-RANDOM SEQUENCES AND FACTORIZATION

The sets $\mathscr{B} = \{b_1, \ldots, b_m\}$ and $\mathscr{A} = \{a_1, a_2, \ldots, a_r\}$ will be taken as non-empty sets of abstract elements with $m > r \geqq 2$. By $\mathscr{B}^\infty$ ($\mathscr{B}^*$, resp.) the set of all infinite (finite, resp.) sequences (strings, words) of elements from $\mathscr{B}$ will be denoted, similarly for $\mathscr{A}^\infty$ and $\mathscr{A}^*$. $\mathscr{B}^* = \bigcup_{n=0}^{\infty} \mathscr{B}^n$, where $\mathscr{B}^0 = \{\Lambda\}$ and Λ is the empty sequence.

Let $U(\mathscr{B})$ ($U(\mathscr{A})$, resp.)be a fixed universal Turing machine over the alphabet $\mathscr{B}$ ($\mathscr{A}$, resp.); the reader is supposed to be familiar with this notion which is defined in [1], [7] or elsewhere. For $p, S, x \in \mathscr{B}^*$, $U(\mathscr{B})(p, S) = x$ means: having written the concatenation $p * S$ on the input tape of $U(\mathscr{B})$ and having initialized it, $U(\mathscr{B})$ eventually terminates its work with x written on the output tape. The *algorithmic complexity* $K_{U(\mathscr{B})}(x \mid S)$ *of* x *given* S *and w.r.t. to* $U(\mathscr{B})$ is defined by

$$(2) \qquad K_{U(\mathscr{B})}(x \mid S) = \min \{n : (\exists p \in B^n)(U(\mathscr{B})(p, S) = x)\} .$$

The set over which the minimalization is taken is always non-empty. In fact, there exists $c = c(U(\mathscr{B})) \in \mathscr{N}$ such that, for all $x, S \in \mathscr{B}^*$, $K_{U(\mathscr{B})}(x \mid S) \leqq l(x) + c$, where $l(x) = n$ iff $x \in \mathscr{B}^n$. As a more detailed introduction into the notion of algorithmic complexity and its properties, in the extend sufficient for our further reasonings, can serve [8] or [2], Chapter 5.

An infinite sequence $x = x_1, x_2, \ldots \in \mathscr{B}^\infty$ is called *absolutely random*, if there exists a total function $f: \mathscr{N} \to \mathscr{N}, f \in o(n)$ (i.e., $\lim_{n \to \infty} f(n)\, n^{-1} = 0$), such that, for all n,

$$(3) \qquad K_{U(\mathscr{B})}(x_1, \ldots, x_n \mid n) \geqq n - f(n) .$$

An immediate generalization of Martin-Löf result ([6], cf. also [2]), shows that absolutely random sequences exist iff, roughly speaking, $\sum_{i=0}^{\infty} r^{-f(i)} < \infty$ and if it is the case, then "almost all" infinite sequences (in the sense of product measure generated by equiprobable distribution on $\mathscr{B}$) satisfy (3).

Given $x \in \mathscr{B}^\infty \cup \mathscr{B}^*$ and $n_1, n_2, \ldots, n_r \in \mathscr{N}$ such that $\sum_{i=1}^{r} n_i = m$, we shall denote by $w(x, n_1, \ldots, n_r)$, or simply by $w(x)$, the sequence $y \in \mathscr{A}^\infty \cup \mathscr{A}^*$ defined by the factorization algorithm described above (i.e., $l(y) = l(x)$, $y_i = a_j$ iff $x_i = b_k$ and $\sum_{s=0}^{j-1} n_s < k \leqq \sum_{s=0}^{j} n_s$, $n_0 = 0$).

For $x \in \mathscr{B}^\infty$, which is absolutely random, the relative frequencies of occurrences of particular letters or blocks of letters tend to equiprobable distribution. In symbols, if $B(k, x) = \langle x_1, \ldots, x_k \rangle, \langle x_{k+1}, \ldots, x_{2k} \rangle, \ldots \in (\mathscr{B}^k)^\infty$, if $B(k, x)[n]$ denotes the initial segment of length n (w.r. to $\mathscr{B}^k$) of $B(k, x)$ and if $fr(z, b)$ denotes the relative frequency of occurrences of b in $z \in \mathscr{B}^*$, then for each $k \in \mathscr{N}^+$ and each $\alpha \in \mathscr{B}^k$,

$$\lim_{n \to \infty} fr(B(k, x)[n], \alpha) = m^{-k} . \tag{4}$$

Hence, if x is absolutely random, then for each $j \leqq r$,

$$\lim_{n \to \infty} fr(w(x)[n], a_j) = n_j/m = p_j , \tag{5}$$

as follows immediately from the factorization algorithm ((4) is proved in [3] or [5]). Of course, $w(x)$ need not be absolutely random sequence w.r. to $\mathscr{A}$ and, in fact, it is not, supposing that $\langle p_1, \ldots, p_r \rangle \neq \langle r^{-1}, r^{-1}, \ldots, r^{-1} \rangle$, as the following construction proves.

Let $y \in \mathscr{A}^*$, denote by $n_j(y)$ the total number of occurrences of a_j in y, so that $\sum_{j=1}^{r} n_j(y) = l(y)$. Let $Q(y) \subset \mathscr{A}^{l(y)}$ be the set of all $l(y)$-tuples with the same numbers of occurrences of particular letters as in y, let $k(y)$ be the order number of y according to the lexicographical ordering generated in $Q(y)$ by the simple ordering $a_1 < a_2 < \ldots$ $\ldots < a_r$ of $\mathscr{A}$. Finally, by $\tilde{p} \in \mathscr{A}^*$ denote the shortest program with this property: for each $y \in \mathscr{A}^*$, and for a fixed encoding of $(r + 1)$-tuples in the form of concatenation,

$$U(\mathscr{A})(\tilde{p} * n_1(y) * n_2(y) * \ldots * n_r(y) * k(y), l(y)) = y \tag{6}$$

(the concatenation in (6) will be denoted by $P_0(y)$). As proved in [5], if $\langle p_1(y), \ldots$ $\ldots, p_r(y) \rangle = \langle n_1(y)(l(y))^{-1}, \ldots, n_r(y)(l(y))^{-1} \rangle \neq \langle r^{-1}, r^{-1}, \ldots, r^{-1} \rangle$, then there exists $c < 1$ and $g \colon \mathscr{N} \to \mathscr{N}$, $g \in o(n)$, such that $l(P_0(y)) \leqq c \, . \, l(y) + g(l(y))$. In fact, we may take $c = H_r(p_1(y), \ldots, p_r(y)) = -\sum_{j=1}^{r} p_j(y) \log_r p_j(y)$.

Hence, if x is an absolutely random sequence from $\mathscr{B}^\infty$, but $\langle n_1, n_2, \ldots, n_r \rangle \neq$ $\neq \langle m/r, m/r, \ldots, m/r \rangle$, then $w(x, n_1, \ldots, n_r) \in \mathscr{A}^\infty$ is not absolutely random as each $w(x)[n]$ can be generated by "substantially shorter" program $P_0(w(x)[n])$ for n large enough. We may ask, whether $w(x)[n]$ can be generated by a still shorter program. The answer is negative, as the following theorem shows.

Theorem 1. Let $\mathscr{A} = \{a_1, \ldots, a_r\}$, $\mathscr{B} = \{b_1, \ldots, b_m\}$, $m > r \geqq 2$, let $U(\mathscr{A})$, $U(\mathscr{B})$ be fixed universal Turing machines over $\mathscr{A}$ and $\mathscr{B}$, let $x \in \mathscr{B}^\infty$ be absolutely random, let $n_1, n_2, \ldots, n_r \in \mathcal{N}, \sum_{i=1}^{r} n_i = m$. Then there exists a total function $g : \mathcal{N} \to \mathcal{N}$, $g(n) \in o(n)$, such that for all $n \in \mathcal{N}$,

$$(7) \qquad l(P_0(w(x, n_1, n_2, \ldots, n_r)[n])) - K_{U(\mathscr{A})}(w(x, n_1, n_2, \ldots, n_r)[n] \mid n) \leqq g(n) .$$

Proof. Let $x = x_1, x_2, \ldots \in \mathscr{B}^\infty$ be absolutely random, so that $K_{U(\mathscr{B})}(x[n] \mid n) \geqq$ $\geqq n - f(n)$ for a function $f(n) \in o(n)$. Let $y = y_1, y_2, \ldots = w(x) \in \mathscr{A}^\infty$ and suppose, in order to arrive at a contradiction, that there exists $\gamma < 1$ such that, for infinitely many n's,

$$(8) \qquad K_{U(\mathscr{A})}(y[n] \mid n) \leqq \gamma \,.\, l(P_0(w(x, n_1, \ldots, n_r)[n])) .$$

A computation yields that the assertion of Theorem 1 will be proved if we arrive at a contradiction supposing that there exists $\beta < 1$ such that, for infinitely many n's,

$$(9) \qquad K_{U(\mathscr{A})}(y[n] \mid n) \leqq \beta H_r(p_1, \ldots, p_r)\, n ,$$

where $p_i = n_i/m, i \leqq r$.

Given x and $n_1, \ldots, n_r$, $w(x, n_1, \ldots, n_r)$ is defined uniquely, but not vice versa. Or, having at our disposal $y = y_1, y_2, \ldots$ with, say, $y_j = a_i$, we need still a number $v_j \leqq n_i$ to know, which b_k was transformed into a_i to be able to reconstruct x_j. Hence, we need a sequence $v = v_1, v_2, \ldots$ of numbers, where $0 < v_i \leqq n_i^*$ for $i =$ $= 1, 2, \ldots$ and $n_i^* = n_j$ iff $y_i = a_j$, namely, we need $y[n]$ and $v[n]$ in order to generate $x[n]$.

Given $y[n]$ the n-tuple $v[n]$ of numbers can be enumerated by a number $\alpha(n)$, $\alpha(n) \leqq \prod_{i=1}^{n} n_i^*$, so that $\log_r (\alpha(n)) \leqq \sum_{i=1}^{n} \log_r n_i^*$. The relative frequences of occurrences of each b_i in x tend to m^{-1}, so that the relative frequency of cases when $n_i^* = n_j$ will tend to n_j/m, or, written precisely, will be $(n_j/m) + g(n)$, where $g(n) \,.\, n \leqq const.$ Hence

$$(10) \qquad \sum_{i=1}^{n} \log_r n_i^* \leqq \sum_{j=1}^{r} n \frac{n_j}{m} \log_r n_j + const = n \sum_{j=1}^{r} \frac{n_j}{m} \log_r n_j + const =$$

$$= n \left[\sum_{j=1}^{r} \frac{n_j}{m} \log_r n_j - \log_r m + \log_r m \right] + const =$$

$$= n \left[\sum_{j=1}^{r} \frac{n_j}{m} \log_r \frac{n_j}{m} + \log_r m \right] + const = n \log_r m - nH_r(p_1, \ldots, p_r) + const,$$

where $p_i = n_i \mid m$. Combining with the assumptions above we obtain

$$(11) \qquad K_{U(\mathscr{A})}(x[n] \mid n) \leqq K_{U(\mathscr{A})}(y[n] \mid n) + K_{U(\mathscr{A})}(v[n] \mid n) + const_1 \leqq$$

$$\leqq \beta H_r(p_1, \ldots, p_r)\, n + n \log_r m - nH_r(p_1, \ldots, p_r) + const_2 =$$

$$= n \log_r m \,.\, \gamma + const_2 ,$$

where

$$\gamma = 1 - (1 - \beta)\, H_r(p_1, \ldots, p_r)\,(\log_r m)^{-1} < 1\,, \tag{12}$$

so that $K_{U(\mathscr{B})}(x[n] \mid n) \leqq n\,.\,\gamma'$ for $\gamma' < 1$ and infinitely many n's, but this contradicts the supposed absolute randomness of x. □

3. PSEUDO-RANDOM INDEPENDENT IDENTICALLY DISTRIBUTED SEQUENCES

Theorem 1 immediately invokes the two following definitions.

Definition 1. A sequence $y \in \mathscr{A}^\infty$ is *strictly pseudo-random independent and identically distributed sequence* (strictly PIID-sequence) with probability distribution $\langle p_1, p_2, \ldots, p_r \rangle$, if there are $n_1, \ldots, n_r \in \mathscr{N}$ such that $n_i/m = p_i$, $i = 1, 2, \ldots, r$, $m = \sum_{i=1}^{r} n_i$, if there is $\mathscr{B} = \{b_1, b_2, \ldots, b_m\}$ and if there is $x \in \mathscr{B}^\infty$ such that x is absolutely random and $y = w(x, n_1, \ldots, n_r)$.

Definition 2. A sequence $y \in \mathscr{A}^\infty$ is PIID-sequence, if there exists a total function $g\colon \mathscr{N} \to \mathscr{N}$, $g \in o(n)$, such that for all $n \in \mathscr{N}$,

$$l(P_0(y[n])) - K_{U(\mathscr{A})}(y[n] \mid n) \leqq g(n)\,. \tag{13}$$

As immediately follows from Theorem 1, each strictly PIID-sequence is a PIID-sequence, but the inverse implication does not hold, as will be proved below.

Theorem 2. Let $x = x_1, x_2, \ldots, \in \mathscr{A}^\infty$ be a PIID-sequence, then for each $a_j \in \mathscr{A}$, $\lim_{n \to \infty} fr(a_j, x[n]) = pr(a_j)$ exists and $\sum_{j=1}^{r} pr(a_j) = 1$.

Proof. Suppose, in order to arrive at a contradiction, that $fr(a_j, x[n])$ oscillates for some $j \leqq r$. Then there exist $\varepsilon > 0$ and an infinite sequence $\bar{n}_1 < \bar{n}_2 < \ldots$ such that $|fr(a_j, x[\bar{n}_i]) - fr(a_j, x[\bar{n}_{i+1}])| > \varepsilon$. Set $n = \bar{n}_i$, $m = \bar{n}_{i+l}$ for some i, l, and consider a program P_1 which, joined with $(r+1)$-tuples $\langle n_1, n_2, \ldots, n_r, k_1 \rangle$, $\langle m_1, m_2, \ldots, m_r, k_2 \rangle$ of integers, calls the program P_0 and generates the k_1th sequence of length $\sum_{s=1}^{r} n_s$ with n_i occurrences of a_i, i.e., the k_1th sequence from $Q(n_1, \ldots, n_r) = Q(a_1^{n_1} \ldots a_r^{n_r})$, then generates the k_2th sequence from $Q(m_1, m_2, \ldots, m_r)$ and joins them in this order. If n_i (m_i, resp.) is the frequency of a_i in $x[n]$ (in $\langle x_{n+1}, \ldots, x_m \rangle$, resp.) and if k_1 (k_2, resp.) is the order number of $x[n]$ ($\langle x_{n+1}, \ldots, x_m \rangle$, resp.) in $Q(n_1, n_2, \ldots, n_r)$ (in $Q(m_1, m_2, \ldots, m_r)$, resp.), then $P_1(x[m]) = P_1 * \langle n_1, \ldots, n_r, k_1 \rangle * \langle m_1, \ldots, m_r, k_2 \rangle$ generates $x[m]$. Clearly,

$$l(P_1(x[m])) \leqq \ulcorner \log_r \|Q(n_1, \ldots, n_r)\| \urcorner + \ulcorner \log_r \|Q(m_1, \ldots, m_r)\| \urcorner + \\ + \sum_{i=1}^{r} \ulcorner \log_r n_i \urcorner + \sum_{i=1}^{r} \ulcorner \log_r m_i \urcorner + l(P_1)\,. \tag{14}$$

es can be easily seen,

$$(15) \qquad \sum_{i=1}^{r} \ulcorner \log_r n_i \urcorner + \sum_{i=1}^{r} \ulcorner \log_r m_i \urcorner + l(P_1) \in o(m)$$

and the assumption that x is a PIID-sequence yields that

$$(16) \qquad |l(P_1(x[m])) - mH_r(p_1, \ldots, p_r)| \in o(m)\,,$$

where $p_i = (n_i + m_i)/m$. The only we have to prove is that there exists $c_0(\varepsilon) < H_r(p_1, \ldots, p_r)$ such that

$$(17) \qquad \log_r \|Q(n_1, \ldots, n_r)\| + \log_r \|Q(m_1, \ldots, m_r)\| \leqq c_0(\varepsilon)\, m\,,$$

where $\|\cdot\|$ denotes the cardinality. But, as proved in [5],

$$(18) \qquad \log_r \|Q(n_1, \ldots, n_r)\| = nH_r(q_1, \ldots, q_r) + f_1(n)\,,$$

$$(19) \qquad \log_r \|Q(m_1, \ldots, m_r)\| = (m - n)\, H_r(s_1, \ldots, s_r) + f_2(m - n)\,,$$

where $q_i = n_i/n$, $s_i = m_i/(m - n)$, $i \leqq r$, and $f_1(n), f_2(n) \in o(n)$. So we must prove that

$$(20) \qquad nH_r(q_1, \ldots, q_r) + (m - n)\, H_r(s_1, \ldots, s_r) \leqq c_0(\varepsilon)\, mH_r(p_1, \ldots, p_r)\,.$$

As $\langle q_1, \ldots, q_r \rangle \neq \langle s_1, \ldots, s_r \rangle$ (recall that $|q_j - s_j| > \varepsilon > 0$), and because of the fact that, for $\alpha = n/m$, $\beta = (m - n)/m$,

$$(21) \qquad \alpha q_i + \beta s_i = (n \mid m)\,(n_1/n) + ((m - n) \mid n)\,(m_i/(m - n)) = \\ = (n_i + m_i)/n = p_i\,,$$

(20) follows from the well-known assertion of classical information theory according to which

$$(22) \qquad \frac{\alpha H_r(q_1, \ldots, q_r) + \beta H_r(s_1, \ldots, s_r)}{H_r(p_1, \ldots, p_r)} < 1\,,$$

so that the left hand side in (22) can be taken as $c_0(\varepsilon)$. $\sum_{j=1}^{r} fr(a_j, x[n]) = 1$ for each $n \in \mathcal{N}$, hence, $\sum_{j=1}^{r} pr(a_j) = 1$ as well. □

Given $\alpha = \langle x_1, \ldots, x_n \rangle \in \mathcal{A}^*$ and $n_1, n_2, \ldots, n_r \in \mathcal{N}$, we may define the *product probability* $\widetilde{pr}(n_1, \ldots, n_r, \alpha)$ of α, setting $\widetilde{pr}(n_1, \ldots, n_r, \alpha) = \prod_{j=1}^{n} \pi(x_i)$, where $\pi(x_i) = p_j = n_j \mid \sum_{k=1}^{r} n_k$ iff $x_i = a_j$. Hence, $\widetilde{pr}(n_1, \ldots, n_r, \alpha)$ is the probability of α supposing that each x_j, $j \leqq n$, is sampled independently from the probability distribution $\langle p_1, \ldots, p_r \rangle$. The assertion of Theorem 2 can be, in the case of strictly PIID-sequences, generalized from particular letters to finite blocks of letters, as the next assertion proves.

Theorem 3. If $y \in \mathscr{A}^\infty$ is a strictly PIID-sequence such that $y = w(x, n_1, n_2, \ldots, n_r)$ for an absolutely random x, $x \in \mathscr{B}^\infty$, $\|\mathscr{B}\| = \sum_{j=1}^{r} n_j$, then for each $k \in \mathscr{N}^+$ and each $\alpha \in \mathscr{A}^k$,

$$\lim_{n\to\infty} fr(\alpha, B(k, y)\,[n]) = \widetilde{pr}(n_1, \ldots, n_r, \alpha)\,. \tag{23}$$

Proof. Let $x = x_1, x_2, \ldots \in \mathscr{B}^\infty$ be an absolutely random sequence over an alphabet $\mathscr{B} = \{b_1, \ldots, b_m\}$, let $n_1, n_2, \ldots, n_r \in \mathscr{N}$, let $\sum_{j=1}^{r} n_j = m$, let $w(x, n_1, n_2, \ldots$ $\ldots, n_r) = y = y_1, y_2, \ldots \in \mathscr{A}^\infty$. For $k, n \in \mathscr{N}^+$ take the nth element of $B(k, y)$, i.e. the k-tuple $\bar{y} = \langle y_{(n-1)k+1}, y_{(n-1)k+2}, \ldots, y_{nk}\rangle \in \mathscr{A}^k$. Its inverse image under the mapping w, i.e. the set

$$C = \{\langle v_1, v_2, \ldots, v_k\rangle : w(\langle v_1, \ldots, v_k\rangle) = \bar{y}\}\,, \tag{24}$$

is a product subset of $\mathscr{B}^k$ such that the projection of C to its jth coordinate has n_i elements iff $y_{(n-1)k+j} = a_i$. Hence, $\|C\| = \prod_{j=1}^{r} n_j^{c_j}$, where c_j is the absolute frequency of a_j in $\bar{y}$, the total cardinality of $\mathscr{B}^k$ is m^k, so that the relative frequency of elements from C in $\mathscr{B}^k$ is $\prod_{j=1}^{r} (n_j/m)^{c_j}$, but this is nothing else than $\prod_{i=1}^{k} \pi(y_{(n-1)k+1}) = \widetilde{pr}(n_1, \ldots$ $\ldots, n_r, \bar{y})$. As x is absolutely random, the same holds for $B(k, x)$, so that the relative frequency of each k-tuple from $\mathscr{B}^k$ in $B(k, x)$ tends to m^{-k}. Consequently, the relative frequency of $\bar{y}$ in $B(k, y)$ tends to $\widetilde{pr}(n_1, \ldots, n_r, \bar{y})$. □

4. PROPERTIES OF PIID-SEQUENCES

A PIID-sequence $y \in \mathscr{A}^\infty$ is called *rational*, if for each $j \leqq r$ the value $\lim_{n\to\infty} fr(a\ , y[n])$ is a rational number. Theorem 3 can be extended to rational PIID-sequences, as the next assertion proves.

Theorem 4. Each rational PIID-sequence $y = y_1, y_2, \ldots \in \mathscr{A}^\infty$ is a strictly PIID-sequence, i.e., there exist a set $\mathscr{B} = \{b_1, \ldots, b_m\}$ and positive integers $n_1, n_2, \ldots, n_r$ such that $\sum_{i=1}^{r} n_i = m$ and $y = w(x, n_1, n_2, \ldots, n_r)$ for an absolutely random $x \in \mathscr{B}^\infty$.

Proof. As proved in Theorem 2, the values $p_j = \lim_{n\to\infty} fr(a_j, y[n])$, $j \leqq r$, exist and are rational, due to the assumptions. Hence, there exist $n_1, \ldots, n_r \in \mathscr{N}$ such that $n_j/m = p_j$, $j \leqq r$, where $m = \sum_{j=1}^{r} n_j$.

Let $\mathscr{B} = \{b_1, b_2, \ldots, b_m\}$ be a set, let $A_1, A_2, \ldots \subset \mathscr{B}$ be non-empty subsets of $\mathscr{B}$ such that $\|A_i\| = n_i^*$ (hence, $n_i^* \leqq m$ for all $i \in \mathscr{N}$, the connection with n_i^*'s introduced above will be seen later), let $S' \in A_0^\infty$, $A_0 \subset B$, be an infinite sequence. First, we prove: if $f(n) \in o(n)$ is a total function such that $\sum_{k=1}^{\infty} m^{-f(k)} < \infty$, then there exists an infinite sequence $S \in \mathop{\mathrm{X}}\limits_{i=1}^{\infty} A_i$ such that, for each $n \in \mathcal{N}$,

$$K_{U(\mathscr{B})}(S[n] \mid S'[n]) > \log_m \Big(\prod_{i=1}^{n} n_i^* \Big) - f(n) . \tag{25}$$

When proving (25) we use some ideas from the proof of Theorem 6, Chapter 5 in [2]. Take $f(n) \in o(n)$ satisfying the conditions (e.g. $f(n) = c \log_m n$, $c > 1$), and set

$$D_n = \Big\{ S \colon S \in \mathop{\mathrm{X}}\limits_{i=1}^{\infty} A_i, K_{U(\mathscr{B})}(S[n] \mid S'[n]) > \log_m \Big(\prod_{i=1}^{n} n_i^* \Big) - f(n) \Big\} . \tag{26}$$

We have to prove that there exists $K \in N$ such that $\bigcap_{n=K}^{\infty} D_n \neq \emptyset$ (replacing $f(n)$ by $f(n) + const$ we may always suppose that $\bigcap_{n=1}^{K-1} D_n = \mathop{\mathrm{X}}\limits_{n=1}^{K-1} A_n$). Let $\mathscr{F}_0 \subset \mathscr{P}(\mathscr{B}^\infty)$ be the minimal σ-field of subsets of $\mathscr{B}^\infty$ generated by elementary cylinders, let $P^\sim$ be the probability measure defined on $\mathscr{F}_0 \cap \mathop{\mathrm{X}}\limits_{i=1}^{\infty} A_i$ as the unique extension of the function $P^\sim$ ascribing to each cylinder $V(x_1, \ldots, x_n) \subset \mathop{\mathrm{X}}\limits_{i=1}^{\infty} A_i$ the value $\prod_{i=1}^{r} (n_i^*)^{-1}$. D_n is a union of cylinders of the length n and its complement w.r.t. $\mathop{\mathrm{X}}\limits_{i=1}^{\infty} A_i$ contains at most

$$\sum_{i=0}^{L} m^i = (m^{L+1} - 1)(m - 1)^{-1} < m^{L+1} = \Big(\prod_{i=1}^{n} n_i^* \Big) m^{1-f(n)} \tag{27}$$

cylinders of the length n, where $L = \log_m \Big(\prod_{i=1}^{n} n_i^* \Big) - f(n)$. Or, programs which generate $S[n]$ using $U(\mathscr{B})$ and $S'[n]$ are sequences from $\mathscr{B}$ and different $S[n]$'s need different programs. The $P^\sim$-probability of each $S[n]$ is $\prod_{i=1}^{n} (n_i^*)^{-1}$, hence,

$$P^\sim \Big(\mathop{\mathrm{X}}\limits_{i=1}^{\infty} A_i - C_n \Big) < m^{1-f(n)} , \tag{28}$$

so that

$$\sum_{i=0}^{\infty} P^\sim \Big(\mathop{\mathrm{X}}\limits_{i=1}^{\infty} A_i - D_n \Big) < \infty ,$$

hence, for some K, $P^\sim \Big(\bigcap_{n=K}^{\infty} D_n \Big) > 0$ and $\bigcap_{n=K}^{\infty} D_n \neq \emptyset$.

Now, let $y = y_1, y_2, \ldots \in \mathscr{A}^\infty$ be a rational PIID-sequence, then each sequence $v \in \mathop{\mathrm{X}}\limits_{i=1}^{\infty} A_i$, where $\|A_i\| = n_i^* = n$ iff $y_i = a_j$, defines uniquely a sequence $x \in \mathscr{B}^\infty$ such that $y = w(x, n_1, \ldots, n_r)$. Hence, there is a fixed program p_1 which, given n, $n_1, \ldots$

$\ldots, n_r, y[n]$ and $v[n]$, computes $x[n]$, another fixed program p_2 computes the concatenation $y[n] * v[n]$ given $n, n_1, \ldots, n_r$ and $x[n]$. So

$$(29) \qquad |K_{U(\mathscr{B})}(x[n] |\, n, n_1, \ldots, n_r) - K_{U(\mathscr{B})}(y[n] * v[n] \,|n, n_1, \ldots, n_r)| < const$$

(different pairs $\langle y[n], v[n]\rangle$ yield different $x[n]$, so $y[n]$ and $v[n]$ can be effectively computed from $x[n], n, n_1, \ldots, n_r$). We may embed $\mathscr{A}$ into $\mathscr{B}$ by identifying, say, each a_j with the first b_k to which a_j corresponds by factorization. So we may take, due to (25), such a $v \in \bigtimes_{i=1}^{\infty} A_i$ that for all $n \in \mathscr{N}$,

$$(30) \qquad K_{U(\mathscr{B})}(v[n] \mid y[n]) > \log_m \left(\prod_{i=1}^{n} n_i^* \right) - f(n)\,,$$

where $f(n)$ is an appropriate $o(n)$-function.

Let q_1 be the shortest program which generates $y[n]$ (in fact, q_1 is $P_0(y[n])$ up to $g(n)$ where $g(n)$ is an $o(n)$-function with respect to which y is a PIID-sequence). So q_1 can be taken as the r-adic number of $y[n]$ in the set of all n-tuples which are possible initial segments of a PIID-sequence with respect to the given $g(n)$, $n, n_1, \ldots, n_r$. Let q_2 be the shortest program which generates $v[n]$ given $y[n], n_1, \ldots$ $\ldots, n_r$, again, q_2 can be taken as the m-adic number of the $2n$-tuple $y[n] * v[n]$ in the set of all $2n$-tuples such that $y[n]$ is given and $v[n] \in \bigtimes_{i=1}^{n} A_i$. Different $y[n] *$ $* v[n]$'s yield different $x[n]$'s, different $\langle q_1, q_2\rangle$'s yield different $y[n] * v[n]$'s, hence, the shortest program for $x[n]$ cannot be shorter than $l(q_1 * q_2) = l(q_1) +$ $+ l(q_2)$ (up to an $o(n)$-function). However, the same argumentation as used in the proof of Theorem 1 shows, that $q_1 * q_2$ cannot be encoded, in $\mathscr{B}$, by a sequence shorter than $n - f'(n)$, if (30) holds, for an appropriate $f'(n) \in o(n)$. Combining this result with (29) we obtain that x is absolutely random and the theorem is proved. □

As the limit probabilities of frequences of letters in strictly PIID-sequences correspond to ratios of n_i and finite cardinality m of the basic alphabet $\mathscr{B}$, only rational PIID-sequences can be obtained by factorization of absolutely random sequences from $\mathscr{B}^\infty$. Let $\langle p_1, \ldots, p_r\rangle$ be any probability distribution including those with irrational p_i's, let C_p be the set of all sequences from $\mathscr{A}^\infty$ which are PIID-sequences with limit values of relative frequencies of letters corresponding to $\langle p_1, \ldots, p_r\rangle$, let $\widetilde{pr}(p_1, \ldots, p_r, \cdot)$ be the probability measure defined in the same way as $\widetilde{pr}(n_1, \ldots$ $\ldots, n_r, \cdot)$, but with n_i/m replaced by p_i. The well-known assertion of information theory and coding theory then enables to prove that $\widetilde{pr}(p_1, \ldots, p_r, C_p) = 1$, hence, the set C_p is non-empty so that there exist non-rational and, consequently, non-strictly PIID sequences.

The following theorems shows that even a slight weakening of the notion of PIID-sequence admits also sequences not satisfying the limit assertion (23) of Theorem 3 for relative frequencies of blocks of letters. Let $c < 1$ be given, a sequence $y \in \mathscr{A}^\infty$

is called a c-PIID-sequence, if for all $n \in \mathscr{N}$

$$K_{U(\mathscr{A})}(x[n] \mid n) \geqq c \,.\, l(P_0(y[n]))\,. \tag{31}$$

Theorem 5. For each $\mathscr{A} = \{a_1, a_2, \ldots, a_r\}$, $r \geqq 2$, each probability distribution $\langle p_1, \ldots, p_r \rangle$, each $c < 1$ and each $\varepsilon > 0$ there exist $\langle p_1^*, \ldots, p_r^* \rangle$ and a c-PIID-sequence $x = x_1, x_2, \ldots, \in \mathscr{A}^\infty$ such that

(i) $p_i^* > 0,\ i \leqq r,\ \sum_{i=1}^{r} p_i^* = 1$,

(ii) $\max\{|p_i - p_i^*| : i \leqq r\} < \varepsilon$,

(iii) $\lim_{n \to \infty} fr(a\ , x[n]) = p_j^*,\ j \leqq r$,

(iv) there exist $k \in \mathscr{N}^+$ and $\alpha \in \mathscr{A}^k$ such that $fr(\alpha, B(k, x)\,[n]) = 0$ for all $n \in \mathscr{N}$ (i.e., α does not occur in $B(k, x)$).

Proof. Evidently, given a probability distribution $\langle p_1, \ldots, p_r \rangle$ and $\varepsilon > 0$, there always exist *positive* integers $n_1, \ldots, n_r \in \mathscr{N}^+$ such that, for $m = \sum_{i=1}^{r} n_i$, $\max\{|p_i - (n_i/m)| :\ i \leqq r\} < \varepsilon$, set $p_i^* = n_i/m$. Consider the sets $\mathscr{B}_k = Q(kn_1, \ldots, kn_r)$, $k \in \mathscr{N}^+$, of all km-tuples with kn_i occurrences of a_i for all $i \leqq r$. Let $y = y_1, y_2, \ldots \ldots \in \mathscr{B}_k^\infty$ be an absolutely random sequence over the new alphabet $\mathscr{B}_k$, let $z = z_1, z_2, \ldots \ldots$ be the sequence y taken as a sequence over the original alphabet $\mathscr{A}$. So $B(km, z) = y$, and each $B(km, z)\,(n) = y_n$ contains at least k occurrences of each a_j (all n_i are positive). Hence, if e.g. $\alpha \in \bigcup_{j=1}^{r} (\mathscr{A} - \{a_j\})^{km}$, then $B(km, z) \neq \alpha$, so that $fr(\alpha, B(km, z)\,[n]) = 0$ for all $n \in \mathscr{N}$. Clearly, $fr(a_j, y_n) = fr(a_j, \langle z_{(n-1)km+1}, \ldots \ldots, z_{nkm} \rangle) = n_i/m = p_i^*$ due to the definition of y_n as a sequence from $Q(kn_1, \ldots, kn_r)$, so that (iii) holds. The only we have to prove is that for each $c < 1$ there exist $k \in \mathscr{N}^+$ such that $z = z(k)$ is a c-PIID-sequence.

The sequence $y \in \mathscr{B}_\kappa^\infty$ is absolutely random, so that for each $\tilde{n}$ and an $o(n)$-function $f(n)$,

$$K_{U(\mathscr{B}_k)}(y[\tilde{n}] \mid \tilde{n}) \geqq \tilde{n} - f(\tilde{n}) \tag{32}$$

with respect to the alphabet $\mathscr{B}_k$. If $q \in \mathscr{B}_k^*$ is the shortest program such that $U(\mathscr{B}_k)\,(q, \tilde{n}) = y[\tilde{n}]$ and $l_{\mathscr{B}_k}(q) = \tilde{n} - f(\tilde{n})$, then q cannot be encoded in the original alphabet $\mathscr{A}$ by a sequence shorter than

$$(\tilde{n} - f(\tilde{n}))\,(\log_r \|B_k\| + g_1(\tilde{n})) + g_2(\tilde{n}) \tag{33}$$

for appropriate $g_1(n), g_2(n) \in o(n)$. But

$$\log_r \|B_k\| = \log_r \|Q(kn_1, \ldots, kn_r)\| = mkH_r(p_1^*, \ldots, p_r^*) - h(mk)\,, \tag{34}$$

$h(n) \in o(n)$, so that, given $c < 1$, we may choose $\mathrm{k} \in \mathscr{N}^+$ such that q cannot be encoded in $\mathscr{B}_k$ by a sequence shorter than $c'(\tilde{n} - f(\tilde{n}))\, H_r(p_1^*, \ldots, p_r^*)$ for a $c' < 1$, but $c' > c$. Given $n \in \mathscr{N}$, the algorithmic complexity of $z[n]$ may differ from the algorithmic complexity of $z[n']$, where $n' \geqq n$ is the smallest integer divisible by km, at most, by a value independent of n ($z_{n+1}, z_{n+2}, \ldots, z_n$, are joined to $z[n]$ and $n' - n < km$). But $z[n'] = y[n'']$ for $n'' = n'/km$. Combining these results we obtain that $z(k)$ is a c'-PIID-sequence, hence, it is also a c-PIID-sequence as $c' > c$. The theorem is proved. □

5. CONCLUSIVE REMARKS

The model explained in this paper is of purely mathematical nature and forms a theoretical background and limitation for models more close to practical use. The theoretical ineffectiveness of the present model goes in two directions:

(1) The notions of absolute random sequence and PIID-sequence are of asymptotic nature, as they are defined up to an $o(n)$-function. Hence, each finite sequence can be an initial segment of an absolute random or PIID-sequence and no definite conclusions about the absolute randomness or PIID-property of infinite sequences can be taken on the grounds of their initial segments. On the other hand, such an asymptotic conception makes the definitions independent of a particular choose of the universal Turing machine U (algorithmic complexities defined w.r. to two universal Turing machines U_1, U_2 differ only by an additive constant which depends only on U_1, U_2, and such a difference is irrelevant from the point of view, how the classes of absolute random sequences and PIID-sequences are defined). The asymptotic feature of this conception may be eliminated by giving a *fixed* universal Turing machine U and a *fixed* $o(n)$-function $f: \mathcal{N} \to \mathcal{N}$, but even in this case it would be interesting to know, which properties of the defined notions are independent of the chosen parameters and it is just what we tried to investigate here. From a point of view the asymptotic notions defined here can be seen as analogies of those statements of classical probability theory which are asymptotically valid "almost surely" or "with the probability one", but it would deserve a more detailed consideration to find how far this analogy goes. Some philosophical remarks concerning the classical "almost surely" valid statements can be found in [2], Chapter 4.

(2) Even with U and f fixed the notions still remain to be ineffective because of the fact that they are defined through an effectively non-computable function $K_U(x| S)$. This difficulty can be partially eliminated when replacing U by a less efficient computational device (by a partial recursive function, formally said) Ψ such that $K_\Psi(x \mid S)$ is computable. An intuitively reasonable example of this approach consists in taking into considerations only such programs p which compute, given S, the desired sequence x within a priori given time and space limitations (in non-trivial cases these limitations may depend on x through, say, $l(x)$, cf. [4]).

On the other hand, the twofold idealization (or abstraction) connected with the model presented here enables to pick out the basic methodological feature of this model which can be easily extended to algorithmical-complexity-based models of more complicated notions of classical probability (e.g. Markov chains). Defining a notion conceived, as it is the rule in classical probability theory, as a property of a generator which produces random outputs, we must always suppose that this property is projected, somehow, into the combinatorial properties of the output sequence (in the opposite case this property would be of purely metaphysical nature with no possibility to test it on the ground of the observed outputs). Now, the combinatorial (algorithmic-complexity based) alternative of the notion in question can be defined by the set of infinite output sequences for which the corresponding combinatorial property represents the shortest way how to define them (the only combinatorial property binding the realizations of classical i.i.d. sequences is the asymptotic stability of relative frequences of particular letters and it is just the property from which the definition of PIID-sequences takes profit). Again, a more detailed formulation and investigation of this methodological principle would deserve some effort.

ACKNOWLEDGEMENT

The author thanks to Jan Šindelář for valuable remarks and comments which much improved the final version of this paper.

(Received September 30, 1985.)

REFERENCES

[1] M. Davis: Computability and Unsolvability. McGraw-Hill New York 1958.

[2] T. L. Fine: Theories of Probability — An Examination of Foundations. Academic Press, New York—London 1973.

[3] I. Kramosil: Monte-Carlo methods from the point of view of algorithmic complexity. In: Trans. 9th Prague Conf. on Inform. Theory, Statist. Dec. Functions and Random Processes, pp. 39—51. Academia, Prague 1983.

[4] I. Kramosil: Recursive classification of pseudo-random sequences. Kybernetika *20* (1984), supplement, pp. 1—34.

[5] I. Kramosil and J. Šindelář: Infinite sequences of high algorithmic complexity. Kybernetika *20* (1984), 6, 459—466.

[6] P. Martin-Löf: The definition of random sequences. Inform. and Control *9* (1966), 602—619.

[7] H. Rogers, Jr.: Theory of Recursive Functions and Effective Computability. McGraw-Hill, New York 1967.

[8] C.-P. Schnorr: Zufälligkeit und Wahrscheinlichkeit. (Lecture Notes in Mathematics 218.) Springer-Verlag, Berlin—Heidelberg—New York 1971.

RNDr. Ivan Kramosil, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Praha 8. Czechoslovakia.