INVARIANT PROPERTIES OF INFORMATIONAL BULKS

L.A. Levin, V.V.V'jugin
U.S.S.R.

1. Introduction

Many properties of informational bulks are not preserved under recodings (for example, from the binary system to a ternary one) and thus they are not properties of the informational bulk itself, but, rather, they characterize the text which is the bearer of this information. Different texts may contain approximately the same information.

Contrary to such properties, our report dwells on invariant properties of informational bulks, i. e. the properties preserved under recodings of the information bearers. For simplicity of mathematical formulation, infinite sequences of natural numbers will be considered as the information bearers (though only finite ones are of practical importance). The recodings are carried out by computable operators: sequences α and β are equivalent, if $\alpha = F(\beta)$ and $\beta = G(\alpha)$ for some computable operators F and G. Such α and β contain approximately the same information up to the finite description of F and G.

Using methods of the theory of algorithms it is possible to prove the existence of sequences which possess quite exotic properties. For example (see [1], § 13.5), there exist sequences containing "indivisible" information: such α is incomputable and for any incomputable β, if $\beta = F(\alpha)$ for some computable operator F, then β is equivalent to α, i. e. the information in α is infinite and equivalent - up to the (finite) description of the recoding algorithm - to any infinite part of it. It is dubious whether sequences with such properties may exist in reality. And in fact, as it was pointed out in [1], in any combination of computable and random processes (in the sense specified below) the probability of obtaining "indivisible" sequences equals 0. The specificness of the present paper is that we eliminate such properties by means of the notion of an "ignorable" set. A set of sequences A is <u>unattainable</u>, if for any computable operator F and any $\omega \notin A$ it holds $F(\omega) \notin A$. A set of sequences B is called <u>ignorable</u>, if it is contained in an unattainable set of the Lebesgue measure 0. Any other computable measure of such a set B equals 0, so the probability of obtaining a sequence from B in any random process with "simple" distribution of probability equals 0. No $\omega \in B$ can be obtained as $\varepsilon = F(\omega')$, where $\omega' \notin B$. So we may claim that no sequence of an ignorable set can be obtained by a physical process, reducible to a combination of algorithmic and random processes. In part 3 of this paper invariant properties are considered "up to the ignorable ones", a Boolean algebra L being introduced for this purpose.

2. Some notions of algorithmic theory of information

Let Ω be a set of all infinite sequences of natural numbers. Concatenating finite sequences α and β, the sequence $\alpha\beta$ can be constructed; $\alpha \subset \beta$, if $\beta = \alpha\delta$ for some δ. ω_i is the i-th member of ω, $(\omega)_i = \omega_1\omega_2 \ldots \omega_i$. $l(\omega)$ is the length of ω. A semimeasure is a function P, defined on the set of all finite sequences, acquiring nonnegative real values and satisfying the condition $P(\Lambda)=1$, $\sum_k P(xk) \leq P(x)$ for all x. A semimeasure P is called a measure, if $\sum_k P(xk) = P(x)$. If P is a measure, then we can define $P\{\omega \in \Omega \mid x \subset \omega\} = P(x)$. Now P may be extended to all measurable subsets of Ω. The Lebesgue measure is a measure L, such that $L(x) = 2^{-l(x)}$, if x contains only zeros and ones, and $L(x)=0$ in the opposite case. Semimeasure P is called recursively enumerable (further r.e.) if $\{(r,x) \mid r < P(x),\ r \text{ is a rational number}\}$ is a r.e. set. If P is a r.e. measure, then $\{(r,x) \mid r > P(x)\}$ is also a r.e. set, so P(x) can be calculated with any degree of accuracy (r.e. measure is also called computable).

Proposition. There exists a r.e. semimeasure M such that for any r.e. semimeasure P a constant C can be found, such that $C.M(x) \geqslant P(x)$ for all x (abbreviated $M(x) \succcurlyeq P(x)$).

Semimeasure M is called universal. With every semimeasure P a maximal measure $\overline{P}$ not exceeding P and satisfying the condition

$$\overline{P}(x) = \inf_n \sum_{x \subset y,\ l(y)=n} P(y)$$

is naturally associated. $\overline{M}$ is called a universal measure. Note that $\overline{M}$ is not a r.e. measure. A computable operator is a r.e. set F consisting of pairs of finite sequences such that if $(x,y) \in F$, $(x',y') \in F$ and $x \subset x'$, then $y \subset y'$.

If $(x,y)\varepsilon F$, we write $y=F(x)$. For $\omega\varepsilon\Omega$, $F(\omega)$ is the union of all sequences $y\subset F((\omega)_n)$. The value of F is defined on $\omega\varepsilon\Omega$ if $F(\omega)\varepsilon\Omega$. It can be shown that any r.e. semimeasure may be presented as $P(x)=L\{\omega|x\subset F(\omega)\}$ for some computable operator F. Thus $\overline{P}(x)=L\{\omega|x\subset F(\omega),\ F(\omega)\varepsilon\Omega\}$.

According to the definition, $A\subseteq\Omega$ is ignorable if and only if $L(F^{-1}(A))=0$ for any computable operator F, i. e. $\overline{P}(A)=0$ for any r.e. semimeasure P. It follows from the above proposition that this is equivalent to the condition $\overline{M}(A)=0$.

If $l(x)=l(y)$, the pair of sequences (x,y) is encoded by the sequence, which we denote as (x,y), such that $(x,y)_i=x_i+\frac{1}{2}(x_i+y_i)\ (x_i+y_i+1)$. On the basis of that we may identify any semimeasure on Ω with the corresponding semimeasure on the set of all pairs of infinite sequences. Let $d(\omega/Q)=\log_2\frac{dM}{dQ}(\omega)$ be deficiency of randomness of ω with respect to Q, where

$$\frac{dP}{dQ}(\omega)=\lim_n\frac{P((\omega_n))}{Q((\omega)_n)}$$

is the Radon-Nicodim derivative of semimeasure P by a semimeasure Q. $P\otimes Q$ is a product of semimeasures P and Q. Let us define the quantity of information in the sequence α about the sequence β as the deficiency of indenpendence of α and β:
$I(\alpha:\beta)=d((\alpha,\beta)/M\otimes M)$.

The following theorem characterizes the class of all ignorable sets in terms of quantity of information.

Theorem 1. A set $A\subseteq\Omega$ is ignorable if and only if there exists $\alpha\varepsilon\Omega$ such that $I(\omega:\alpha)=\infty$ for any $\omega\varepsilon A$.

The value $|\log_2 M(x)|$ characterizes the full quantity of information in the finite sequence x (the complexity of x). Unfortunately it is incomputable. We shall single out the class of sequences not having this defect. The sequence $\omega\varepsilon\Omega$ is called complete, if $M((\omega)_n)\asymp P((\omega)_n)$ for some r.e. measure P. ($\Phi\asymp F$ means $\Phi\preccurlyeq F$ and $F\preccurlyeq\Phi$). Such ω contains all information necessary for the computation of its complexity by means of P. It can be shown that for any r.e. measure P, $M((\omega)_n)\asymp P((\omega)_n)$. if and only if ω is random with respect to the measure P in the sense of Martin Löf [2].

Theorem 2. Any r.e. measure of the set of all complete sequences equals 1 and any everywhere defined computable operator transforms it into itself.

This theorem shows that only complete sequences can be obtained in any combination of random processes with "simple" distribution of probability and everywhere defined algorithmic processes.

3. Algebra L and its natural elements

The set $A \subseteq \Omega$ is invariant, if with each sequence it contains all equivalent sequences. Let K be the class of all invariant subsets Ω measurable by the universal measure $\overline{M}$ (specifically, K contains all invariant Borel subsets of Ω). K is the Boolean algebra with the usual set-theoretic operations. Let us define the equivalence relation on K : $A \approx B$, if $(A \setminus B) \cup (B \setminus A)$ is ignorable. Taking the quotient of K modulo this equivalence relation we obtain a Boolean algebra L. The meaning of these notions can be clarified in the following way. Each invariant property of informational bulks defines the invariant set of sequences. If two such properties define the sets, which differ from each other on an ignorable set, then, according to the introduction of this paper, we can claim that they define the same set of natural informational bulks (i. e. a set that can be obtained in some combinations of algorithmic and random natural processes). Such elements of L correspond to the properties of real informational bulks. Each specific mathematical property of sequences can be defined by the formula of a strictly determined language. The requirement of measurability by measure $\overline{M}$ is much weaker; L may contain elements undeterminable by any mathematical properties. But all elements of L constructed further will be defined by some formulae of the arithmetical language.

The simplest elements of L : 0 is defined to be the class of all ignorable sets, 1 is defined to be the class containing Ω. These are the minimal and the maximal elements of L.

Let us consider the elements formed by complete sequences, and sequences equivalent to them. The information contained in a complete sequence can be encoded with "maximal density". One should be aware of the fact that any information in the densest codification becomes indistinguishable by its properties from random noise and, therefore, the only invariant characteristic of such information is its quantity (finite or infinite). This fact is illustrated by Theorem 3. Each computable sequence is complete. Let C be the set of all computable sequences, $c \in L$ be such that $C \in c$. R is the set of all incomputable complete sequences, and sequences equivalent to them, $\mathfrak{r} \in L$ is such that $R \in \mathfrak{r}$ It is evident that c is an atom of L, i. e. $c \neq 0$ and it cannot be represented as $c=a\cup b$ where $a\cap b=0$, $a\neq 0$, $b\neq 0$.

Theorem 3. τ is an atom of L and, thus, any element of L, formed by complete sequences (and by those equivalent to them), coincides with 0, c, τ or $c\cup\tau$.

In *[3]* these results are applied to the problems of intuitionism. It is shown below that in a commom case the situation is more complex since not all sequences allow optimal encoding.

4. Elements of L, formed by sequences not equivalent to complete ones

The main result of this part is that the set $\Omega\setminus(C\cup R)$ is not ignorable (see *[4]* for proof), moreover, sequences of this set form an infinite number of various elements of L. The information contained in any sequence from $\Omega\setminus(C\cup R)$ does not allow optimal encoding. $p=1\setminus(c\cup\tau)$.

Theorem 4. The set of all atoms of L is countable.

For each atom x distinct from c and τ, $x\subseteq p$ i. e. x is formed by sequences from $\Omega\setminus(C\cup R)$.

Sequences of the atom-forming set are not distinguishable by any invariant property. However among the sequences from $\Omega\setminus(C\cup R)$ there are ones possessing infinitely divisible properties. Let $d_0, d_1, \ldots d_n, \ldots$ be all atoms of L.

Theorem 5. $d=1\setminus\bigcup_{i=0}^{\infty} d_i\neq 0$.

The element d has the property that for any $x\subseteq d$, x is not an atom.

We shall point out in conclusion, that it follows from Theorem 2, that algorithmic processes which are not universally defined are essential for the obtaining of sequences which are non-equivalent to complete ones.

The main results of Sections 2 and 3 are due to L. Levin, the results of Section 4 are due to V.V. V'jugin.

BIBLIOGRAPHY

1. Rogers, H., Theory of recursive functions and effective computability, McGraw-Hill, New York, 1967.
2. Martin-Löf, P., Definition of random sequences, Information and Control, 9, (1966) N6, 602-619.
3. Levin, L.A., On the principle of conservation of information in intuitionistic mathematics, Dokl. Akad. Nauk SSSR, 227 (1976) N6, 1293-1296.
4. V'jugin, V.V., On Turing invariant sets, Dokl. Akad. Nauk SSSR, 229, (1976) N4, 790-793.