Coding-invariant Classification of Infinite Sequences

Vladimir V. V’yugin

Institute for Information Transmission Problems
Russian Academy of Sciences

CIRM, Marseille 29 June - 3 July 2009

Coding invariant properties

Ω - set of all infinite binary sequences, where $\alpha, \beta \in \Omega$;

$\alpha \prec \beta$ if $\alpha = F(\beta)$ for some computable operation F (Turing reducible);

$\alpha \equiv \beta$ (Turing equivalent) if $\alpha \prec \beta$ and $\beta \prec \alpha$.

Informally speaking, the sequences $\alpha \equiv \beta$ have the same information content (up to finite descriptions of the coding methods).

$A \subseteq \Omega$ - coding invariant if $\omega \in A$ and $\alpha \equiv \omega$ implies $\alpha \in A$ for any $\omega, \alpha \in \Omega$ (A consists of Turing degrees).

Apriory semimeasure

Ξ - set of all finitesequences, $\mathscr{R}^+$ - set of all nonnegative real numbers;

A function $P : \Xi \to \mathscr{R}^+$ is r.e. semimeasure if
$P(\Lambda) = 1$;
$P(x0) + P(x1) \leq P(x)$;
$\{(r, x) : r \in \mathscr{Q}, r < P(x)\}$ is r.e. set, where r is a rational number.

M - the apriory semimeasure: for any r.e. semimeasure P a constant c exists such that $M(x) \geq cP(x)$ for all $x \in \Xi$.

Extension on the set Ω

$\mathscr{B}$ - Borel sigma-algebra on Ω generated by intervals

$\Gamma_x = \{\omega \in \Omega : x \subseteq \omega\}$

P - a semimeasure,
$\bar{P}$ - the maximal measure such that $\bar{P} \leq P$:

$$\bar{P}(\Gamma_x) = \inf_n \sum_{x \subseteq y, l(y)=n} P(y)$$

$\bar{P}$ is extended on all Borel subsets of Ω

Negligible sets

A is negligible if $\bar{M}(A) = 0$.

A is negligible if and only if

$\bar{P}(A) = 0$ for each r.e. semimeasure P;

$L\{\alpha : F(\alpha) \in A\} = 0$ for each computable F, [1]

$\bar{M}\{\omega : \exists\alpha(\alpha \in A \& \alpha \prec \omega)\} = \bar{M}(\cup_F F^{-1}(A)) = 0$.

---

[1] since $\bar{P}(A) = L(F^{-1}(A))$, where L is uniform Lebesgue measure on Ω, $L(\Gamma_x) = 2^{-l(x)}$.

Algebra of invariant properties

$A \subseteq \Omega$ is coding invariant if $\omega \in A$ and $\alpha \equiv \omega$ implies $\alpha \in A$ for any $\omega, \alpha \in \Omega$.

I - Boolean algebra of all coding invariant Borel subsets of Ω;

$\Upsilon = I/\sim$ - factor-algebra of I with respect to the equivalence relation

$$A \sim B \Longleftrightarrow \bar{M}((A \setminus B) \cup (B \setminus A)) = 0,$$

where $A, B \in I$.

Trivial elements

$\mathbf{a} = [A]$ - the element of Υ defined by a coding invariant set $A \in I$.

$P(\mathbf{a}) = \bar{P}(A)$ for any semicomputable semimeasure P, where $A \in \mathbf{a}$.

$\mathbf{0} = [\emptyset]$ - zero element - consists of all coding invariant subsets of Ω of $\bar{M}$-measure 0, $\bar{M}(\mathbf{0}) = 0$,

$\mathbf{1} = [\Omega]$ - unit element, $\bar{M}(\mathbf{1}) > 0$,

Random sequences

By Zvonkin and Levin for any sequence ω Martin-Löf random with respect to a computable measure

- ω is computable, or
- $\omega \equiv \alpha$, where α is Martin-Löf random with respect to the uniform measure L.

$km(\alpha^n) = n + O(1)$, where α^n is the prefix of α of length n, and $km(x)$ is monotonic complexity;
α can be optimally compressed.

Elements of Υ generated by random sequences:

$\mathbf{c} = [C]$, where C - set of all computable sequences, $\bar{M}(\mathbf{c}) > 0$,

$\mathbf{r} = [\bar{R}]$, where R - set of all sequences random with respect to the uniform measure L and $\bar{R} = \{\omega : \exists\alpha(\alpha \in R \& \alpha \equiv \omega)\}$,

$\bar{M}(\mathbf{r}) > 0$ and $P(\mathbf{r}) = 0$ for each computable measure P.

$\mathbf{r}$ contains information which can be optimally compressed.

$\mathbf{d} \in \Upsilon$ - atom if $\mathbf{d} \neq \mathbf{0}$ and it cannot be represented as $\mathbf{d} = \mathbf{a} \cup \mathbf{b}$, where $\mathbf{a} \cap \mathbf{b} = \emptyset$, $\mathbf{a} \neq \mathbf{0}$ and $\mathbf{b} \neq \mathbf{0}$.

This theorem is a corollary of Kolmogorov's $0-1$ -law.

Theorem

(Leonid Levin).
The elements $\mathbf{r}$ *and* $\mathbf{c}$ *are atom elements of* Υ*.*

Optimally compressed nontrivial information cannot be distinguished from a random noise.

Sketch of the proof.

$(\Omega, L, \mathscr{B})$ - probability space, where
$\mathscr{B}$ - Borel field generated by all random variables $f_n(\omega) = \omega_n$,
$n = 1, 2, \ldots$, $\omega = \omega_1 \omega_2 \ldots$ is a binary sequence.

$f_1, f_2, \ldots$ - independent random variables.

$\mathscr{B}_n = \mathscr{B}(f_n, f_{n+1}, \ldots)$ - Borel field generated by $f_n, f_{n+1}, \ldots$.

$A \in \mathscr{B}_i$ for all i, where A is invariant with respect to inverting of a finite number of bits.

Since $A \in \cap_i \mathscr{B}_i$, by Kolmogorov's $0-1$ - law $L(A) = 0$ or $L(A) = 1$.

R - the set of all Martin-Löf random with respect to L sequences is invariant with respect to inverting of a finite number of bits.

$A \subset R$ and $\bar{M}(A) > 0$ then $L(A) > 0$.

If $R = A \cup B$, where A, B - invariant Borel sets then $L(A) = 0$ or $L(B) = 0$.

Then $\bar{M}(A) = 0$ or $\bar{M}(B) = 0$,

since $\bar{L}(A) = \int_A \frac{dL}{d\bar{M}}(\omega) d\bar{M}$ and $\frac{dL}{d\bar{M}}(\omega) > 0$ for $\bar{M}$-almost all $\omega \in R$.

Any set of sequences generating an atom cannot be divided on two parts of positive measure $\bar{M}$ by their information content.

A problem arises (Leonid Levin):

Does $\mathbf{1} = \mathbf{r} \cup \mathbf{c}$?

Informally,

Does any infinite non-computable sequence can be optimally compressed?

Theorem

$\mathbf{d} = \mathbf{1} \setminus (\mathbf{r} \cup \mathbf{c}) \neq \mathbf{0}$.

By this theorem $M(D) > 0$, where $\mathbf{d} = [D]$.

Equivalently, $L\{\omega : F(\omega) \in D\} > 1 - \varepsilon$ for some computable operation F, where $\varepsilon > 0$.

We can generate on probabilistic machine sequences which cannot be optimally compressed.

In particular, these sequences cannot be Turing equivalent to sequences Martin-Löf random with respect to a computable measure.

Main tool

$\Omega_P = \Omega \setminus \cup_{P(x)=0}\Gamma_x$ - support set of a semimeasure P;
Evidently, $\bar{P}(\Omega) = \bar{P}(\Omega_P)$.

A semimeasure P (and the corresponding measure $\bar{P}$) is non-degenerate if $\bar{P}(\Omega) > 0$.

A non-degenerate measure $\bar{P}$ is concentrated on its support set Ω_P.

Application to algorithmic randomness

Proposition

There exist a non-degenerate r.e. semimeasure P such that for any $\omega \in \Omega_P$ and for any computable operation F, $F(\omega)$ is not Martin-Löf random with respect to the uniform measure L.

Sketch of the proof of Proposition.

Directed graph:

Ξ - the set of all vertices, G - the set of edges, where

$(x, x0)$, $(x, x1)$ are directed edges of unit length;

(x, y) - extra edges, where $x, y \in \Xi$, $x \subset y$ and $l(y) - l(x) \geq 2$.

Network: on this graph with one source λ (empty sequence) and with no sink:

$q(x, y)$ - network capacity of the edge $(x, y) \in G$:

$$\sum_{y:(x,y)\in G} q(x, y) \leq 1$$

and $q(x, x0) = q(x, x1)$ for all $x \in \Xi$.

Frame function:

$$R(\lambda) = 1;$$
$$R(y) = \sum_{(x,y)\in G} q(x,y)R(x)$$

for $y \neq \lambda$.

Network flow (with delays) is the minimal semimeasure $P \geq R$.

By definition P is r.e semimeasure.

$\bar{P}(\Omega_P)$ - flow value.

The goal of the construction:

$F_i(x)$ - a sequence of all computable operations;

Each extra edge σ corresponds to some task $\langle i, k\rangle$, where i is the program of F_i and k is the number of a session.

The goal of the task $\langle i, k\rangle$: to define a finite set of extra edges such that for each infinite binary sequence ω:

- either ω contains some extra edge (x, y) as a subword $x \subset y \subset \omega$ such that $l(F_i(y)) > n(x, k)$, where $n(x, k)$ is the order number of the pair (x, k), or
- $s(x) = 1$ for some $x \subset \omega$, where $s(x)$ -a network flow delay function, if no such extra edge (x, y) exists; $s(x) = 1 - q(x, x0) - q(x, x1)$.

$q(x, y)$ determines a portion of the network flow $P(x)$ outgoing from the vertex x along the extra edge (x, y);

$s(x) = \Delta$ for all x of some length n, where Δ is sufficiently small; a portion of the network flow which is temporary delayed;

$q(x, x0) = q(x, x1) = \frac{1}{2}(1 - s(x))$ for these x;

Δ is released if $d_i(y)$ sufficiently large for some y such that $x \subset y$; In this case:

- define $q(x, y) = s(x) = \Delta$ and add the extra edge (x, y);
- delay a new Δ-portion of network flow on all y such that $x \not\subset y$.

More details:

- If no such y exists then $F_i(\omega)$ is finite; only a small portion Δ of the network flow is delayed on all steps.
- If such y exists, the Δ-portion of the network flow is directed along the extra edge (x, y). We define $s(z) = \Delta/(1 - \Delta))$ for all z of length $l(y)$ such that $x \subset z$ and $z \neq y$. Again, a portion Δ of the network flow is delayed.

Since $s(\omega^n)$ increases along each infinite ω, each session proceeds at a finite number of steps.

If $F_i(\omega)$ is infinite then for any k, $F_i(\omega)$ belongs to the union of a r.e. sequence of intervals defined by extra edges (x, y) of the measure $\sum_{(x,y)} 2^{-n(x,k)} \leq 2^{-ck}$.

The information-theoretic structure of the algebra Υ is described by the following theorem.

Theorem

Let $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \ldots$ *be all atom elements of* Υ*. The decomposition*

$$\mathbf{1} = \cup_{i=1}^{\infty} \mathbf{a}_i \cup \mathbf{d}$$

of the unit element of the algebra Υ *is valid , where*

- *the atom* $\mathbf{a_1} = \mathbf{r}$ *is generated by all Martin-Löf random sequences;*
- *the atom* $\mathbf{a}_2 = \mathbf{c}$ *is generated by all computable sequences;*
- *the atoms* $\mathbf{a}_3, \mathbf{a}_4, \ldots$ *are generated by infinite sequences which cannot be information equivalent to Martin-Löf random sequences;*
- $\mathbf{d}$ *is the non-zero infinitely divisible element of* Υ *generated by the measure-theoretic complement of all atoms.*

By definition the element $\mathbf{d} = \mathbf{1} \setminus \bigcup_{i=1}^{\infty} \mathbf{a}_i$ is infinitely divisible, i.e., for any non-zero $\mathbf{x} \subseteq \mathbf{d}$ a decomposition $\mathbf{x} = \mathbf{x}_1 \cup \mathbf{x}_2$ is valid, where $\mathbf{x}_1 \cap \mathbf{x}_2 = \mathbf{0}$, $\mathbf{x}_1 \neq \mathbf{0}$ and $\mathbf{x}_2 \neq \mathbf{0}$.

Details of the proof

$(\Omega, \bar{P}_m, \mathscr{B})$ - probability spaces, $m = 3, 4, \ldots$.

$\mathbf{a}_m = [U_m]$, where

$U_m = \{\omega \in \Omega_{P_m} : \frac{d\bar{P}_m}{d\bar{M}}(\omega) > 0\}$ (Radon - Nicodim derivative);

$c\bar{M}(U_m) > P_m(\mathbf{a}_m) = \bar{P}_m(U_m) = \bar{P}_m(\Omega_{P_m}) > 0$

$\mathbf{a}_m$ is an atom

By construction there exists a sequence $n_1 < n_2 < \dots$ such that

$$f_i(\omega) = \omega_{n_i+1}\omega_{n_i+2}\dots\omega_{n_{i+1}}, i = 1, 2, \dots$$

is a sequence of independent random variables, where $\omega \in \Omega$;

$\mathscr{B}_i = \mathscr{B}(f_i, f_{i+1}, \dots)$ - descending sequence of Borel fields.

$A \in \mathscr{B}_i$ for all i for any Borel A invariant with respect to inverting of a finite number of bits.

Since $A \in \cap_i \mathscr{B}_i$, by Kolmogorov's $0 - 1$ - law $\bar{P}_m(A) = 0$ or $\bar{P}_m(A) = \bar{P}_m(\Omega) \Rightarrow \bar{M}(A) = 0$ or $\bar{M}(A) = \bar{M}(\Omega_{P_m})$.

Related results

$d : \Xi \rightarrow \mathscr{R}^+$ be a computable function.

$d(\omega) = sup_n d(\omega^n)$ - extension on infinite sequences ω.

Proposition

There exist a non-degenerate r.e. semimeasure P such that for any $\omega \in \Omega_P$

$$d(\omega) = \infty$$

for each computable function d such that $d(\alpha)$ is unbounded on each interval $\Gamma_{\omega^n} = \{\alpha \in \Omega : \omega^n \subset \alpha\}$, $n = 1, 2, \ldots$.

Corollary

There exist a non-degenerate r.e. semimeasure P such that each $\omega \in \Omega_P$

- *is not random with respect to any effectively compact class of (non-computable) probability measures (Bernoulli measures, Markov chains of finite order, stationary measures);*
- *the pointwise ergodic theorem is not valid for ω for each (maybe non-computable) stationary measure P for each computable real function f and computable transformation T preserving the measure P:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(\omega))$$

does not exists.

Sketch of the proof of Corollary.

- the sequence $d_i(x)$ contains all universal tests for the classes of measures;
- the sequence $d_i(x)$ contains all lower semicomputable Bishop's tests $d(x)$ (defined in terms of computable f and T) such that

$$d(\omega) < \infty \Rightarrow \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(\omega)) \text{ exists.}$$