Theoretical Computer Science 588 (2015) 2–15

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

On Martin-Löf (non-)convergence of Solomonoff's universal mixture

Tor Lattimore *, Marcus Hutter

RSCS, Australian National University, Australia

ARTICLE INFO

Article history:
Received 30 September 2013
Received in revised form 17 November 2014
Accepted 6 December 2014
Available online 16 December 2014

Keywords:
Solomonoff induction
Kolmogorov complexity
Theory of computation

ABSTRACT

We study the convergence of Solomonoff's universal mixture on individual Martin-Löf random sequences. A new result is presented extending the work of Hutter and Muchnik [3] by showing that there does not exist a universal mixture that converges on all Martin-Löf random sequences. We show that this is not an artifact of the fact that the universal mixture is not a proper measure and that the normalised universal mixture also fails to converge on all Martin-Löf random sequences.

1. Introduction

Sequence prediction is the task of predicting symbol α_n having seen $\alpha_{1:n-1} = \alpha_1 \cdots \alpha_{n-1}$. Solomonoff approached this problem by taking a Bayesian mixture over all lower semicomputable semimeasures where complex semimeasures were assigned lower prior probability than simple ones.[1] He then showed that, with probability one, the predictive mixture converges (fast) to the truth for any computable measure [9]. Solomonoff induction arguably solves the sequence prediction problem and has numerous attractive properties, both technical [9,2,5] and philosophical [8]. There is, however, some hidden unpleasantness, which we explore in this paper.

Martin-Löf randomness is the usual characterisation of the randomness of individual sequences [6]. A sequence is Martin-Löf random if it passes all effective tests, such as the laws of large numbers and the iterated logarithm. Intuitively, a sequence is Martin-Löf random with respect to measure μ if it satisfies all the properties one would expect of an infinite sequence sampled from μ. It has previously been conjectured that the set of Martin-Löf random sequences is precisely, or contained within, the set on which the Bayesian mixture converges.

This question has seen a number of attempts with a partial negative solution and a more detailed history of the problem by Hutter and Muchnik [3]. They showed that there exists a universal lower semicomputable semimeasure M and Martin-Löf random sequence α (with respect to the Lebesgue measure λ) for which $M(\alpha_n|\alpha_{<n}) \not\to \lambda(\alpha_n|\alpha_{<n})$. The α used in their proof is computable from the halting problem, which presumably inspired the work in [7] where it is shown that if α is 2-random, then every universal lower semicomputable semimeasure converges on α. It is worth remarking that there exist semimeasures that do converge on all Martin-Löf random sequences, some of which are even lower semicomputable.

* Principal corresponding author.
E-mail addresses: tor.lattimore@anu.edu.au (T. Lattimore), marcus.hutter@anu.edu.au (M. Hutter).

[1] Actually, Solomonoff mixed over proper measures. The use of semimeasures was introduced later by Levin to ensure that the mixture itself was lower semicomputable [14].

http://dx.doi.org/10.1016/j.tcs.2014.12.004
0304-3975/

Unfortunately, however, they are not universal and may not enjoy the same fast convergence rates in expectation as universal measures do. For the construction and a detailed discussion, see [4, §5].

While Hutter and Muchnik showed that there exists a universal lower semicomputable semimeasure and Martin-Löf random sequence on which it fails to converge, the question of whether or not this failure occurs for all such semimeasures has remained open. We prove that for every universal lower semicomputable Bayesian mixture there exists a Martin-Löf random sequence on which it fails to converge. This result is interesting for a few reasons. The choice of universal mixture is akin to choosing an optimal universal Turing machine when computing Kolmogorov complexity. In both cases, asymptotic results are rarely dependent on this choice and so it is useful to confirm this trend here. On the other hand, if the result had been positive then the existence of a universal mixture that did converge on all Martin-Löf random strings would be a nice property that might justify the choice of one universal mixture over another.

The universal mixture is not a proper measure in the sense that the sum of conditional probabilities $M(0|x) + M(1|x) < 1$ for all x. For this reason it is common to use a normalised version M_{norm} where normalisation is chosen to preserve the ratio $M_{\text{norm}}(x0)/M_{\text{norm}}(x1) = M(x0)/M(x1)$. We show that the situation is not improved by normalisation and that M_{norm} also fails to converge to the Lebesgue measure on some Martin-Löf random sequences.

Our paper is structured as follows. We present the required notation and some basic results in algorithmic information theory (Section 2). We then present Solomonoff's original theorem showing that the universal mixture converges to the truth with probability one (Section 3). The main theorems are then presented of which Theorem 6 is the most important stating for any universal mixture M that there exists a Martin-Löf random sequence α such that the predictive distribution $M(\alpha_n|\alpha_{<n})$ does not converge to $\frac{1}{2}$ and actually is bounded away from $\frac{1}{2}$ for a non-zero fraction of the time (Section 4). We then show that this is also true of the normalised version of the universal mixture (Section 5) and that there exists an infinite sequence that is not Martin-Löf random, but on which all universal mixtures converge to $\frac{1}{2}$ (Section 6). We conclude in Section 7.

2. Notation

Overviews of algorithmic information theory can be found in [5,1]. A table of notation may be found in Appendix B.

General The natural, rational and real numbers are denoted by $\mathbb{N}$, $\mathbb{Q}$ and $\mathbb{R}$. Logarithms are taken with base 2. A real $\theta \in (0, 1)$ has entropy $H(\theta) := -\theta \log \theta - (1-\theta)\log(1-\theta)$. The indicator function is $[\![expr]\!]$, which takes value 1 if *expr* is true and 0 otherwise. For sets A and B we write $A - B$ for their difference and $|A|$ for the size of A and $A^c = \mathbb{N} - A$ for the complement of A. The empty set is denoted by $\emptyset$. If $A \subseteq \mathbb{N}$ and $n \in \mathbb{N}$, then $A[n] := \{a \in A : a \leq n\}$. We use $\vee$ and $\wedge$ for logical or/and respectively.

Natural density Let $A \subseteq B \subseteq \mathbb{N}$. Then the (upper) natural density of $A \subseteq B$ are

$$d(A, B) := \lim_{n\to\infty} \frac{|A[n]|}{|B[n]|} \qquad \bar{d}(A, B) := \limsup_{n\to\infty} \frac{|A[n]|}{|B[n]|}$$

where the latter quantity is useful in the case when the former does not exist. If $B = \mathbb{N}$, then we abbreviate $d(A) \equiv d(A, \mathbb{N})$ and $\bar{d}(A) \equiv \bar{d}(A, \mathbb{N})$.

Strings A finite binary string x is a finite sequence $x_1x_2x_3 \cdots x_n$ with $x_i \in \mathcal{B} := \{0, 1\}$. Its length is $\ell(x)$. An infinite binary string ω is an infinite sequence $\omega_1\omega_2\omega_3 \cdots$. The empty string of length zero is denoted by ϵ (distinct from $\varepsilon > 0 \in \mathbb{R}$). The sets $\mathcal{B}^n$, $\mathcal{B}^*$ and $\mathcal{B}^\infty$ are the sets of all strings of length n, all finite strings and all infinite strings respectively. Substrings of $x \in \mathcal{B}^* \cup \mathcal{B}^\infty$ are denoted by $x_{s:t} := x_s x_{s+1} \cdots x_{t-1} x_t$ where $s, t \in \mathbb{N}$ and $s \leq t$. If $s > t$, then $x_{s:t} := \epsilon$. A useful shorthand is $x_{<t} := x_{1:t-1}$. Let $x, y \in \mathcal{B}^*$, then $\#x(y)$ is the number of (possibly overlapping and wrapping around) occurrences of x in y and xy is their concatenation. For example, $\#010(1010) = 2$ (because we count the wrap around match when starting at the last bit). If $\ell(y) \geq \ell(x)$ and $x_{1:\ell(x)} = y_{1:\ell(x)}$, then we write $x \sqsubseteq y$ and say x is a prefix of y. Otherwise we write $x \not\sqsubseteq y$. A string $\omega \in \mathcal{B}^\infty$ is normal if $\forall x \in \mathcal{B}^*, \lim_{n\to\infty} \#x(\omega_{1:n})/n = 2^{-\ell(x)}$.

Measures and semimeasures A semimeasure is a function $\mu : \mathcal{B}^* \to [0, 1]$ satisfying $\mu(\epsilon) \leq 1$ and $\mu(x) \geq \mu(x0) + \mu(x1)$ for all $x \in \mathcal{B}^*$. It is a measure if both inequalities are replaced by equalities. A function $\mu : \mathcal{B}^* \to \mathbb{R}$ is lower semicomputable if the set $\{(x, r) : r < \mu(x), r \in \mathbb{Q}, x \in \mathcal{B}^*\}$ is recursively enumerable. In this case there exists a recursive sequence $\mu_1, \mu_2, \cdots$ of computable functions approximating μ from below. For $b \in \mathcal{B}$ and $x \in \mathcal{B}^*$, $\mu(b|x) := \mu(xb)/\mu(x)$ is the μ-probability that x is followed by b. The Lebesgue measure is $\lambda(x) := 2^{-\ell(x)}$.

Complexity A Turing machine T is a recursively enumerable set of pairs of binary strings $T := \{(p^1, x^1), (p^2, x^2), \ldots\}$ where the program p^k outputs x^k. It is a prefix machine if the set of programs is prefix free, $p^k \not\sqsubseteq p^j$ for all $j \neq k$. T is a monotone machine if $p^k \sqsubseteq p^j \implies x^k \sqsubseteq x^j \vee x^j \sqsubseteq x^k$. For prefix machine T the prefix complexity with respect to T is a function $K_T : \mathcal{B}^* \to \mathbb{N}$ defined by

$$K_T(x) := \min_p \left\{ \ell(p) : (p, x) \in T \right\}$$