The Review of Symbolic Logic
Volume 15, Number 3, September 2022

A NOTE ON THE LEARNING-THEORETIC CHARACTERIZATIONS OF RANDOMNESS AND CONVERGENCE

TOMASZ STEIFER
Polish Academy of Sciences

Abstract. Recently, a connection has been established between two branches of computability theory, namely between algorithmic randomness and algorithmic learning theory. Learning-theoretical characterizations of several notions of randomness were discovered. We study such characterizations based on the asymptotic density of positive answers. In particular, this note provides a new learning-theoretic definition of weak 2-randomness, solving the problem posed by (Zaffora Blando, Rev. Symb. Log. 2019). The note also highlights the close connection between these characterizations and the problem of convergence on random sequences.

§1. Introduction. The notion of randomness is at the very core of fundamental ideas of philosophy and science. As such it comes with its own package of puzzles and enigmas. For example, suppose you are faced with some experimental data and you want to learn about the underlying phenomenon—is it deterministic (say, we observe the infinite sequence of zeros $0, 0, \ldots$) or is it random (e.g., the outcomes of a fair and unbiased coin tossing)? Does it even make any sense to say that an individual object (i.e., an infinite sequence of bits) is random?

Computability theory gives us some tools to deal with this problem. For example, we could say that the sequence is random if we cannot predict it well enough using any effective procedure or we could argue that the random sequences are exactly those that are incompressible. Algorithmic randomness theory studies various answers formulated exactly from that point of view. It is now one of the most active and fruitful branches of modern computability theory, drawing attention of researchers from mathematical logic, as well as from the foundations of probability theory. The cornerstone of this theory is the notion of randomness proposed in 1966 by Martin-Löf [15]. Roughly speaking, a sequence is random in the Martin-Löf sense if it does not have any effectively rare property, i.e., property of measure zero that could be tested in a sufficiently effective way. Here, the *effectiveness* is explicated by means of computability. Since then, many other notions of randomness were introduced and studied, constituting an infinite hierarchy of concepts. Furthermore, it was soon observed that the same notions of randomness may be characterized using independent paradigms such as compressibility and betting strategies.

As we have already noted, computability theory provides some perspective on what is an effective procedure and what is not. As it happens, the notion of effectiveness

Received: April 15, 2020.
2020 *Mathematics Subject Classification*: 03D32.
Key words and phrases: algorithmic randomness, learning theory, effectivization.

doi:10.1017/S1755020321000125

is relevant to other areas of philosophical and scientific investigations. Consider the problem of learning. Can we apply the perspective of computability here as well? Gold [11] and Putnam [21] thought so. Again, let us see an example. We task a student (often called an agent or a learner) with a learning problem such as: are all dogs green? We supply the student with data and examples, in this case, examples of dogs. Such data may be represented mathematically, e.g., by an infinite binary sequences (one means a green dog, zero means a different color). Each time a new example is given, the student makes a guess—yes or no. One of the answers is the correct one, the one we want to hear. We could expect the student to stop making mistakes at some point. Is there a computable method which, if followed, leads to such outcome? Now, the dogs are easy (say yes as long as you see only green dogs) but of course, it is not hard to come up with more difficult tasks. This framework—called algorithmic learning theory—may serve as a model for various scenarios, e.g., binary classification problems or a choice of a true physical theory. Sometimes, it may be impossible to stop making mistakes at all. In such case, a liberal teacher may come up with weaker criteria of success (such as giving the correct answer infinitely many times).

Another task that fits well in the learning-theoretic framework is that of detection of rare properties, i.e., deciding whether a given binary sequence belongs to some set of measure zero. This is basically something we would expect of randomness, namely, that a set of random outcomes does not have any rare properties that could be recognized in an effective way. This connection between algorithmic randomness and algorithmic learning was explored by Osherson and Weinstein [18]. They provided new characterizations of the classes of weakly 1-random and weakly 2-random sequences, both of which are readily interpretable in terms of learning and recognition. In a more recent work, Zaffora Blando [27] described slightly more involved characterizations of Martin-Löf randomness and Schnorr randomness.

All these definitions may be interpreted in the following manner—a sequence is algorithmically random if and only if no computable agent recognizes the sequence as possessing some rare property. The difference between these characterizations boils down to what criterion of success is assumed. For example, a sequence x is weakly 1-random if and only if there is no computable agent which gives the negative answer infinitely many times with probability one, yet they give the negative answer only finitely many times on prefixes of x.

This note consists of two parts. In the first, I investigate some criteria of success based on the asymptotic density of affirmative answers, answering the question asked by Zaffora Blando [27]. On the way, I show novel criteria corresponding to the notions of weak 1-randomness and weak 2-randomness.

In the second part, I argue that learning-theoretic characterizations of randomness may be reinterpreted in terms the effectivization of probabilistic theorems (and vice versa). Suppose we have defined a notion of randomness with respect to some computable measure μ, which will be called the class of μ-random sequences. Such class is of μ-measure one. Now, let μ be a computable probability measure on infinite sequences. In modern probability theory, many results are stated in the following form

$$\mu(\{\omega : \phi(\omega)\}) = 1,$$

where ϕ is some formula—often stating a pointwise convergence. The above is usually stated as "$\phi(\omega)$ for μ-almost every ω." In computable measure theory, we seek for

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

effective versions of such theorems, that is we want to know if

$$\phi(\omega) \text{ for every } \mu\text{-random } \omega.$$

Roughly speaking, the difference between non-effective and effective theorems is somewhat similar to the difference between sentences "all cats but one is black" and "Fluffy is the only non-black cat." You may try to formulate effective theorems for various different notions of randomness. Due to historical reasons, much of the attention was given to Martin-Löf randomness. We already know effective versions of many textbook results, e.g., the law of iterated logarithm [26], Doob's martingale convergence theorem [24] or even Birkhoff's ergodic theorem [3, 9, 25]. In some cases the standard proofs are already constructive and the effectivization follows after simple modifications but it is not always the case. Moreover, negative results also exist. For instance, in the context of Solomonoff induction [22, 23], Lattimore and Hutter [14] discovered that no universal mixture (of lower semicomputable semimeasures) converges on all Martin-Löf random sequences. This result motivated Milovanov [16] to find a new universal induction method which does converge on all Martin-Löf random sequences. Moreover, mathematicians also gave some attention to effectivization with respect to Schnorr randomness (e.g., [10, 19]).

It is actually a folklore result that there exists a computable sequence of functions converging almost surely which fails to converge on some Martin-Löf random sequence. We can interpret this fact in the learning-theoretic framework. At the same time, we can straightforwardly translate the results formulated in the learning-theoretic context into statements about convergence on random sequences. My attention here will focus on the convergence in Cesàro averages.

§2. Preliminaries. Before moving to the main results, we introduce some notational conventions and provide some preliminary definitions. The set of all finite words over the binary alphabet $\{0, 1\}$ is denoted by $2^{<\mathbb{N}}$, while the set of all one-sided infinite sequences is denoted by $2^{\mathbb{N}}$. By convention, bits are indexed from 0. Given a word or a sequence x, we let x_i denote the $(i+1)$-th bit and, given $i < j$, we let x_i^j denote a subword $x_i x_{i+1} \dots x_j$ of x consisting of all the digits of x from x_i to x_j. The empty set is denoted by $\Box$. Given a word w, $|w|$ stands for its length. We write $x \preceq y$ to say that x is a prefix of y. We use $\#(A)$ to denote the cardinality of a set A.

2.1. Effective reals. *Computably enumerable* is abbreviated as *c.e.* A real $r > 0$ is called computable (lower semicomputable) if the left cut of r, i.e., $\{q \in \mathbb{Q} : q < r\}$, is computable (c.e.). A function $f : 2^{<\mathbb{N}} \to \mathbb{R}$ is called computable (lower semicomputable) if its values are uniformly computable (lower semicomputable) in $2^{<\mathbb{N}}$. A real is upper semicomputable if its negation is lower semicomputable. In a similar manner, we can define Δ_2^0 reals as those for which the left cut is a Δ_2^0 set.

These reals have a natural characterization in terms of densities. We say that a sequence x has the density r if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} x_i = r.$$

We say that the density of x is undefined if such limit does not exist.

As it happens, Δ_2^0 reals are exactly the densities of computable sequences.

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

THEOREM 2.1 [12]. *A real is Δ^0_2 if and only if is the density of a computable sequence.*

We can go further and define computablity of functions from infinite sequences into reals. A function $f : 2^{\mathbb{N}} \to \mathbb{R}$ is computable if there is a Turing functional Φ which given oracle x computes the left cut of $f(x)$.

2.2. Probability measures. We are dealing with the binary stochastic process $X = X_0, X_1, \ldots$. Symbol X is introduced to compress notation. For instance, given some formula ϕ and a measure μ we will often write $\mu(\phi(X))$ instead of $\mu(\{x \in 2^{\mathbb{N}} : \phi(x)\})$. X obeys similar notational conventions as sequences, e.g., X^i_j denotes random variables $X_j, X_{j+1}, \ldots, X_i$ and so on. A special attention is given to the uniform measure λ on the Cantor space of infinite binary sequences. This measure corresponds to $\lambda(\{x \in 2^{\mathbb{N}} : x_0^{|\sigma|-1} = \sigma\}) = 2^{-|\sigma|}$ for all nonempty $\sigma \in 2^{<\mathbb{N}}$. Given a word $\sigma \in 2^{<\mathbb{N}}$ we define the cylinder set $[\![\sigma]\!]$ as the set $\{x \in 2^{\mathbb{N}} : x_0^{|\sigma|-1} = \sigma\}$. Similarly, if V is a set of words, then $[\![V]\!] = \cup_{\sigma \in V} [\![\sigma]\!]$. Unless it is stated otherwise, μ denotes an arbitrary computable probability measure, i.e., a probability measure such that there exists a computable function $f : 2^{<\mathbb{N}} \times \mathbb{N} \to \mathbb{Q}$ with $|f(\sigma, n) - \mu([\![\sigma]\!])| < 2^{-n}$.

A measure μ is called continuous if $\mu(\{x\}) > 0$ for no $x \in 2^{\mathbb{N}}$. The reader is encouraged to consult [5] for an introduction to modern measure-theoretic probability theory.

2.3. Learnable sequences. From the learning-theoretic perspective, Δ^0_2 sequences from the arithmetical hierarchy are of a special interest. These sequences are sometimes called *learnable*—this name is justified by the following theorems.

THEOREM 2.2 [11], [21]. *A sequence $x \in 2^{\mathbb{N}}$ is Δ^0_2 iff there exists a total computable $g : \mathbb{N}^2 \to \{0, 1\}$ such that for all $i \in \mathbb{N}$ we have*

$$\lim_{t \to \infty} g(i, t) = x_i.$$

We can interpret function g as a learner which makes guesses about the true value of x_i. A sequence is learnable if as some point, the answers stabilize on a correct one. However, we might also give a slightly different learning-theoretic characterization. In the second scenario, which will be explained in detail in Section 2.5, the learner reads fragments of a sequence and tries to find out whether the sequence has some property.

PROPOSITION 2.3 (FOLKLORE?). *For every $x \in 2^{\mathbb{N}}$ which is Δ^0_2, there exists a computable function $f : 2^{<\mathbb{N}} \to \{0, 1\}$ such that x is the only sequence for which $\#\{i : f(x^i_0) = 1\}$ is infinite.*

Proof. Let $x \in 2^{\mathbb{N}}$ be Δ^0_2. By Theorem 2.2 there exists a computable function $g : \mathbb{N}^2 \to \{0, 1\}$ such that for all $i \in \mathbb{N}$ we have $\lim_{t \to \infty} g(i, t) = 1$ iff $x_i = 1$ and $\lim_{t \to \infty} g(i, t) = 0$ iff $x_i = 0$. We define f by induction. We also define an auxiliary function u. Let $f(\square) = 0$ and $u(\square) = 0$. Suppose that for some σ we have already defined $f(\sigma)$ and $u(\sigma)$ and we want to define $f(\sigma b)$ (with $b \in \{0, 1\}$). Consider a sequence $w = g(0, |\sigma|)g(1, |\sigma|) \ldots g(|\sigma|, |\sigma|)$. Let k be the length of the longest prefix of w which is also a prefix of σb. If $k > u(\sigma)$, let $f(\sigma b) = 1$. Otherwise, we let $f(\sigma b) = 0$. Finally, set $u(\sigma b) = \max(\{k, u(\sigma)\})$. It remains to observe that for each n, there exists m such that $g(a, b) = x_a$ for all $a < n$ and $b > m$. Moreover, there is no $y \neq x$ such that this happens. Hence, f answers 1 on infinitely many prefixes of x and on only finitely many prefixes of every other sequence. □

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

2.4. Martin-Löf randomness. Several equivalent definitions of Martin-Löf randomness—referred to as 1-randomness here—are now known. We start with the definition by tests. A reader interested in learning more about the algorithmic randomness theory is referred to [8].

DEFINITION 2.4. A collection $U_0, U_1, \ldots$ of sets of sequences is uniformly c.e. if and only if there is a collection $V_0, V_1, \ldots \subset 2^{<\mathbb{N}}$ such that $U_i = [\![V_i]\!]$ for every $i \in \mathbb{N}$ and $V_0, V_1, \ldots$ are uniformly c.e.

DEFINITION 2.5 (Martin-Löf μ-test). A uniformly c.e. sequence $U_0, U_1, \ldots$ of sets of sequences is called a Martin-Löf μ-test if there exists a computable f such that $\lim_{n\to\infty} f(n) = 0$ and $\mu(U_n) \leq f(n)$ for every $n \in \mathbb{N}$.

DEFINITION 2.6 (Martin-Löf μ-randomness). A sequence $x \in 2^{\mathbb{N}}$ is called 1-random with respect to μ (or 1-μ-random) if there is no Martin-Löf μ-test $U_0, U_1, \ldots$ such that $x \in \cap_{i\in\mathbb{N}} U_n$.

When dealing with sequences random with respect to some arbitrary computable measure μ, we will usually refer to these simply as 1-*random sequences*.

The following is a folklore result.

PROPOSITION 2.7 (FOLKLORE). *There exists a* Δ^0_2 λ*-random sequence.*

As in case of arithmetic hierarchy, we can define a hierarchy of complexities of classes. A set $C \subseteq 2^{\mathbb{N}}$ is called a Σ^0_n class if there exists a computable relation R such that for all $x \in 2^{\mathbb{N}}$ we have $x \in C$ if and only if $\exists i_1 \forall i_2 \ldots \exists i_n R(x_0^{i_1}, x_0^{i_2}, \ldots, x_0^{i_n})$ for an odd n and $\exists i_1 \forall i_2 \ldots \forall i_n R(x_0^{i_1}, x_0^{i_2}, \ldots, x_0^{i_n})$ for even n. Now, it is possible to give the definition of weak n-randomness.

DEFINITION 2.8 (WEAK n-RANDOMNESS). A sequence is called weakly n-μ-random if it is contained in every Σ^0_n class of μ-measure one.

As in the case of 1-randomness, if we are dealing with an arbitrary computable measure μ we omit μ when referring to weak n-μ-randomness.

2.5. Learning and randomness. A learning-theoretic characterization of two notions of weak n-randomness was discovered by Osherson and Weinstein. The function f in the following statements formalizes the notion of a computable agent—also called a learner—who tries to learn from the prefixes of the sequence whether the sequence possess some rare property or not. The agent is reading bits of a sequence and gives a positive or a negative answer after reading each bit. A positive answer is interpreted as a sign of belief that the given sequence manifest a certain pattern or property we want to detect. The procedure is constrained by the requirement of computability.

It is assumed that a purely random sequence should not have any non-trivial rare properties that could be detected by such learner. It is now a question of the choice of a criterion of success for such an agent. One idea is to ask for only finitely many negative answers. By Theorem 2.9 this criterion may be used to define the class of weakly 1-random sequences.

THEOREM 2.9 (Osherson-Weinstein [18]). *A sequence* x *is weakly* 1*-random if and only if there is no computable function* $f : 2^{<\mathbb{N}} \to \{0, 1\}$ *such that*

$$\#\{i : f(x_0^i) = 0\} < \infty$$

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

and

$$\mu(\#\{i : f(X_0^i) = 0\} < \infty) = 0.$$

A weaker criterion—by Theorem 2.10 corresponding to weak 2-randomness—is given by the requirement of infinitely many positive answers.

THEOREM 2.10 (Osherson-Weinstein [18]). *A sequence x is weakly 2-random if and only if there is no computable function $f : 2^{<\mathbb{N}} \to \{0, 1\}$ such that*

$$\#\{i : f(x_0^i) = 1\} = \infty$$

and

$$\mu(\#\{i : f(X_0^i) = 1\} = \infty) = 0.$$

Finally, a recent theorem by Zaffora Blando [27] gives a learning-theoretic characterization of Martin-Löf randomness.

THEOREM 2.11 (Zaffora Blando [27]). *A sequence x is 1-random if and only if there is no computable function $f : 2^{<\mathbb{N}} \to \{0, 1\}$ such that*

$$\#\{i : f(x_0^i) = 1\} = \infty$$

and for all $n \in \mathbb{N}$

$$\mu(\#\{i : f(X_0^i) = 1\} > n) \leq 2^{-n}.$$

§3. Density of answers. The learning-theoretic characterizations of Martin-Löf randomness and Schnorr randomness were obtained by Zaffora Blando [27] by taking a notion of success present in the Osherson–Weinstein characterization of weak 2-randomness (infinitely many positive answers) and tweaking the measure-theoretic condition in the definition. Naturally, one may wonder, whether similar goal could be obtained by tweaking the success notion instead. To this end, Zaffora Blando [27] asked about notions of randomness arising when we enrich the learning-theoretic approach with conditions imposed on the density of positive answers. In particular, she formulated the following problem.

PROBLEM 3.12. *Consider a class $\mathcal{LD}$ of all sequences $x \in 2^{\mathbb{N}}$ such that there is no computable function g satisfying both of the following*:

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n} g(x_0^i)}{n+1} = 1$$

and

$$\mu\left(\lim_{n\to\infty} \frac{\sum_{i=0}^{n} g(X_0^i)}{n+1} = 1\right) = 0.$$

Does $\mathcal{LD}$ correspond to any known notion of algorithmic randomness?

Such notion of success has a natural interpretation, namely, that we allow the learner to make mistakes but if we look at the average answer, it approaches one. In other words, as time passes, the frequency of negative answers becomes negligible.

An immediate corollary of Theorems 2.9 and 2.10 is that $\mathcal{LD}$ is located between weak 2-randomness and weak 1-randomness. We are going to strengthen this by proving that $\mathcal{LD}$ is, in fact, equal to weak 2-randomness.

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

THEOREM 3.13. *A sequence x is weakly 2-random if and only if there is no computable function $g : 2^{<\mathbb{N}} \to \{0,1\}$ such that*

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n} g(x_0^i)}{n+1} = 1$$

and

$$\mu\left(\lim_{n\to\infty} \frac{\sum_{i=0}^{n} g(X_0^i)}{n+1} = 1\right) = 0.$$

Proof. ($\Leftarrow$) We prove this implication by contraposition. Suppose that x is not weakly 2-random. By Theorem 2.10, there is a computable function $f : 2^{<\mathbb{N}} \to \{0,1\}$ such that

$$\#\{i : f(x_0^i) = 1\} = \infty$$

and

$$\mu(\#\{i : f(X_0^i) = 1\} = \infty) = 0.$$

We will construct a computable function $g : 2^{<\mathbb{N}} \to \{0,1\}$. Let $g(\square) = 1$. Suppose that for some w we have already defined $g(v)$ for all $v \preceq w$ but $g(\tau)$ is yet not defined on any τ—a proper extension of w. Let k be the number of times f gives the positive answer on some prefix of w, i.e.,

$$k = \#\{v : v \preceq w \wedge f(v) = 1\}.$$

For all $\tau \in 2^i$, where $i \leq k$ we let $g(w\tau) = 1$. This completes the construction of the computable function g. Now, observe that for every $y \in 2^{\mathbb{N}}$

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n} g(X_0^{i+1})}{n+1} = 1,$$

if and only if there are infinitely many n such that $f(y_0^n) = 1$. In fact, if there are exactly k indexes n such that $f(y_0^n) = 1$, then

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n} g(X_0^{i+1})}{n+1} = \frac{k}{k+1}.$$

Finally, we may also conclude that

$$\mu\left(\lim_{n\to\infty} \frac{\sum_{i=0}^{n} g(X_0^i)}{n+1} = 1\right) = 0.$$

($\Rightarrow$) This implication follows immediately from Theorem 2.10. □

Now, for completeness we observe that the similar characterization—based on the density of positive answers—may be given for weak 1-randomness.

THEOREM 3.14. *A sequence $x \in 2^{\mathbb{N}}$ is weakly 1-random if and only if there is no computable function $g : 2^{<\mathbb{N}} \to \{0,1\}$ such that*

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n} g(x_0^i)}{n+1} = 1.$$

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

and

$$\mu\left(\lim_{n\to\infty}\frac{\sum_{i=0}^{n} g(X_0^i)}{n+1}=0\right)=1.$$

Proof. ($\Rightarrow$) Fix $x\in 2^{\mathbb{N}}$. Suppose that there is a computable function $g:2^{<\mathbb{N}}\to\{0,1\}$ such that

$$\mu\left(\lim_{n\to\infty}\frac{\sum_{i=0}^{n} g(X_0^i)}{n+1}=0\right)=1$$

and

$$\lim_{n\to\infty}\frac{\sum_{i=0}^{n} g(x_0^i)}{n+1}=1.$$

Consider $f:2^{<\mathbb{N}}\to\{0,1\}$ such that for every $w\in 2^{<\mathbb{N}}$ we let $f(w)=1$ if and only if $\frac{\sum_{i=0}^{|w|-1} g(w_0^i)}{|w|}>1/2$. Observe that if the ratio of the positive answers given by g converges to 0 on some sequence, then f gives the negative answer on infinitely many prefixes. This happens with probability 1. On the other hand, we have $\#\{i: f(x_0^i)=0\}<\infty$. Consequently, x is not weakly 1-random.

($\Leftarrow$) Suppose that x is not weakly 1-random. By Theorem 2.9, there exists a computable function $f:2^{<\mathbb{N}}\to\{0,1\}$ such that

$$\#\{i: f(x_0^i)=0\}<\infty$$

and

$$\mu(\#\{i: f(X_0^i)=0\}=\infty)=1.$$

Let $k=\#\{i: f(x_0^i)=0\}$. We construct a computable function g. Let $g(\square)=0$ and suppose that for some $w\in 2^{<\mathbb{N}}$ we have already defined $g(v)$ for all proper prefixes v of w and we want to define $g(w)$. Let $m=\#\{i<|w|: f(w_0^i)=0\}$. We let $g(w)=1$ if and only if $m\le k$. Otherwise, set $g(w)=0$.

Observe that for every sequence y if there are no more than k indexes i such that $f(y_0^i)=0$ then $g(y_0^j)=1$ for all indexes j. This is true for $y=x$.

On the other hand, $g(X_0^j)=0$ for all but finitely many indexes j when $\#\{i: f(X_0^i)=0\}=\infty$. This happens almost surely. □

As it happens, the values placed as the limits in the last theorem (i.e., one and zero) may be substituted for arbitrary Δ_2^0 reals. Note that this result does not seem to have a straightforward learning-theoretic interpretation and is given as a technical curiosity.

THEOREM 3.15. *Let a and b be Δ_2^0 reals (with $a\neq b$). A sequence $x\in 2^{\mathbb{N}}$ is weakly 1-random if and only if there is no computable function $g:2^{<\mathbb{N}}\to\{0,1\}$ such that*

$$\lim_{n\to\infty}\frac{\sum_{i=0}^{n} g(x_0^i)}{n+1}=a.$$

and

$$\mu\left(\lim_{n\to\infty}\frac{\sum_{i=0}^{n} g(X_0^i)}{n+1}=b\right)=1.$$

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

Proof. Fix two Δ^0_2 reals a and b (with $a \neq b$). By Theorem 2.1 there exist computable $x, y \in 2^{\mathbb{N}}$ such that a is the density of x and b is the density of y. Let f be a computable function witnessing that a sequence ω is not weakly 1-random (in the sense of Theorem 2.9). Let $k = \#\{i : f(x^i_0) = 0\}$. We construct a computable function g. Let $g(\Box) = x_0$ and suppose that for some $w \in 2^{<\mathbb{N}}$ we have already defined $g(v)$ for all proper prefixes v of w and we want to define $g(w)$. Let $m = \#\{i \leq |w| : f(w^i_0) = 0\}$. Check if $m \leq k$. If so, let $g(w) = x_{|w|}$. Otherwise, set $g(w) = y_{|w|}$.

Observe that for every sequence ω if there are no more than k indexes i such that $f(\omega^i_0) = 0$ then the density of $g(\omega^0_0)g(\omega^1_0)\ldots$ equals the density of x, i.e., it is equal to a. This happens if $\omega = x$. Otherwise, this density equals b. This happens with probability one. The implication in the other direction is analogous to the one in the proof of Theorem 3.14. □

§4. Convergence on random sequences. Combining Propositions 2.3 and 2.7 gives the following folklore result as a corollary.

PROPOSITION 4.16 (FOLKLORE). *There exists a* 1*-*λ*-random sequence* x *and a computable function* $f : 2^{<\mathbb{N}} \to \{0, 1\}$ *such that* x *is the only sequence for which* $\#\{i : f(x^i_0) = 1\}$ *is infinite.*

In other words, there are random sequences which are uniquely recognizable by a computable agent, in a certain relaxed sense.

With that in mind, we turn our attention to the problem of convergence on random sequences mentioned in the introduction. We are interested in doing statistical inference based on a finite but increasing amount of data, i.e., we want to study functions which take prefixes of increasing length and output estimates of some parameter (e.g., the entropy rate). This may involve such tasks as hypothesis testing or inductive learning in the form of estimation of the conditional probabilities.

Think of a computable function g which converges to some random variable Y almost surely and on every random sequence, i.e., for every random sequence x we have

$$\lim_{n\to\infty} g(x^n_0) = Y(x).$$

Take the function f from Proposition 4.16 and consider a function h defined by $h(w) = f(w) + g(w)$ for all $w \in 2^{<\mathbb{N}}$. It follows that h converges to Y almost surely but it fails to converge to Y on some λ-random sequence. On the other hand, if Y is computable, then it is a folklore observation that the convergence of g to Y on every weakly 2-random sequence follows from the convergence with probability one. Indeed, we have:

PROPOSITION 4.17 (FOLKLORE). *Let* $g : 2^{<\mathbb{N}} \to \mathbb{R}^{\geq 0}$ *be a computable function such that* μ*-almost surely*

$$\lim_{n\to\infty} g(X^n_0) = Y.$$

If Y *is a computable random variable, then this happens on every weakly* 2*-random sequence* x*.*

Proof. This is a simple consequence of the fact that avoiding a computable limit with an error bounded from below by a rational is a Π^0_2 property. To be precise, for

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

every $i \in \mathbb{N}$ the following is a Π^0_2 class:

$$\{x \in 2^{\mathbb{N}} : \forall n \exists m > n |g(x_0^m) - Y(x)| > 2^{-i}\}.$$

By the assumption this is a class of measure zero and so, no weakly 2-random belongs to it. □

Now, suppose we have two computable functions $h_1, h_2 : 2^{<\mathbb{N}} \to \mathbb{Q}^{\geq 0}$ such that almost surely $\lim_{n\to\infty} h_1(X_0^n) = \lim_{n\to\infty} h_2(X_0^n)$. Such a pair corresponds to an infinite family of computable learning functions $f_1, f_2, \ldots$ defined by

$$\forall (i \in \mathbb{N}) f_n(z_0^i) = 1 \iff |h_1(z_0^i) - h_2(z_0^i)| > 2^{-n}.$$

where $z \in 2^{\mathbb{N}}$ is arbitrary. What can be immediately observed, each such function gives only finitely many positive answers on a weakly 2-random sequence (by Theorem 2.10).

Furthermore, Theorem 2.11 may be reinterpreted in the following form.

THEOREM 4.18. *A sequence $x \in 2^{\mathbb{N}}$ is 1-random if and only if for every $m \in \mathbb{N}$ and any pair of computable functions $h_1, h_2 : 2^{<\mathbb{N}} \to \mathbb{R}^{\geq 0}$ satisfying for all $n \in \mathbb{N}$*

$$\mu(\#\{i \in \mathbb{N} : |h_1(X_0^i) - h_2(X_0^i))| > 2^{-m}\} \geq n) \leq 2^{-n},$$

we have

$$\lim_{n\to\infty} h_1(x_0^n) = \lim_{n\to\infty} h_2(x_0^n)).$$

Proof. For the first implication, simply observe that given $m \in \mathbb{N}$ the sequence $U_1, U_2, \ldots$ defined by

$$U_n = \{z \in 2^{\mathbb{N}} : \#\{i \in \mathbb{N} : |h_1(z_0^i) - h_2(z_0^i))| > 2^{-m}\} \geq n\}$$

is a μ-test. If a sequence x is such that it is not true that

$$\lim_{n\to\infty} h_1(x_0^n) = \lim_{n\to\infty} h_2(x_0^n)$$

then for sufficiently large m we have $x \in \bigcap_{n>0} U_n$, so x is not 1-random.

The second implication follows directly from Theorem 2.11. If y is not 1-random then there is a learning function f which witnesses this. Now, setting $h_1(\sigma) = f(\sigma)$ and $h_2(\sigma) = 0$ for all $\sigma \in 2^{<\mathbb{N}}$ gives what is needed. □

In a way, these two interpretative frameworks, i.e., the detection of rare properties and convergence of estimators, are closely connected. On the one hand, take an appropriate learning function and add it to an estimator. That procedure will render it bad on some random sequences. On the other hand, take a pair of estimators, monitor their difference and you will get a learning function. In such case, the asymptotic behavior of the learning function imitates that of the estimators.

So far, we have considered pointwise convergence of the estimators. This a relatively strong property. Indeed, many inductive schemes do not satisfy pointwise convergence and are optimal only in terms of some weaker criterion of success. Specifically, mathematicians and statisticians studied, with great attention, the convergence in Cesàro averages. Given a function $f : 2^{<\mathbb{N}} \to \{0, 1\}$ we say that f converges to Y

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

on x in Cesàro averages if

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n} f(x_0^i)}{n+1} = Y(x).$$

Such form of convergence is very natural and well studied in summability theory (cf. [20]). There are plenty of natural examples of infinite sequences of reals which do not converge pointwise, but converge in Cesàro averages. In statistics, this may be pictured by the following scenario. We want to estimate a property of the underlying process (such as the entropy rate). As new data comes, we make a new estimation. It is often assumed that more data means a better estimate but it is not always the case. Suppose that an unlikely (but of a positive measure) outcome causes a large error in the estimation. This will happen rarely (as the event in question is of small probability) but nevertheless, it will happen infinitely often. In many cases, such problem may be alleviated by simply averaging all the estimates made so far and using the average as a new estimator. For instance, consider a problem of forward conditional measure estimation for stationary ergodic processes. It was shown by Bailey [2] that the pointwise estimators do not exist in this case but there are known estimators which converge almost surely in Cesàro averages.

Unsurprisingly, there are computable estimators which converge in Cesàro averages to some random variable μ-almost surely but fail to do so on some random point. This prompts a question—under what conditions is convergence (pointwise or in Cesàro averages) on all 1-random sequences guaranteed? In particular, we might be interested in conditions stated in purely probabilistic terms. A partial answer to this is given by the effective version of Breiman's ergodic theorem. We state it in a specialized form below but, firstly, an additional comment is required. For a binary alphabet, a measure μ is stationary if $\mu(X_i = 1)$ is constant for all i. By the Kolmogorov extension theorem (cf. [5]), a stationary measure on the space of sequences from $2^{\mathbb{N}}$ may be uniquely extended to a measure on the space of two-sided infinite sequences (elements of $2^{\mathbb{Z}}$). Similarly, the canonical process $X_0, X_1, \ldots$ is uniquely extended to the process $\ldots X_{-1}, X_0, X_1, \ldots$

THEOREM 4.19. *Let $g : 2^{<\mathbb{N}} \to \mathbb{R}^+$ be a computable function with $\lim_{n\to\infty} g(X_0^n)$ existing almost surely and $\mathbb{E}_\lambda(\sup_i |g(X_0^i)|) < \infty$. Then for every λ-random sequence $\omega \in 2^{\mathbb{N}}$,*

$$\limsup_{k\to\infty} \frac{1}{k+1} \sum_{i=0}^{k} g(\omega_0^i) \le \mathbb{E}_\lambda(\limsup_{k\to\infty} g(X_{-k}^{-1}))$$

and

$$\liminf_{k\to\infty} \frac{1}{k+1} \sum_{i=0}^{k} g(\omega_0^i) \ge \mathbb{E}_\lambda(\liminf_{k\to\infty} g(X_{-k}^{-1})).$$

Proof. The result follows from the effective Birkhoff's ergodic theorem [3, 9, 25] using the proof of Breiman [6]. For details see, e.g., [7]. □

Note that the uniform measure λ in Theorem 4.19 may be substituted by an arbitrary stationary ergodic computable measure. To keep the presentation simple, we choose not to introduce this class of measures in detail here. The curious reader is referred to [5].

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

While learning-theoretic definitions of [18] and later of [27] correspond to the problem of pointwise convergence, the density based characterizations of the type discussed in this work are easily interpreted in terms of convergence in Cesàro averages. To this end, we show yet another learning-theoretic characterization of weak 2-randomness. Here, we consider learning functions with the asymptotic frequency of positive answers equal to zero, almost surely. The theorem states that if the average of initial answers does not converge to zero on some sequence then this sequence is not weakly 2-random.

THEOREM 4.20. *The sequence $x \in 2^{\mathbb{N}}$ is weakly 2-random if and only if there is no computable function f such that*

$$\limsup_{n\to\infty} \frac{\sum_{i=0}^{n} f(x_0^i)}{n+1} > 0,$$

while

$$\mu\left(\lim_{n\to\infty} \frac{\sum_{i=0}^{n} f(X_0^i)}{n+1} = 0\right) = 1.$$

Proof. ($\Leftarrow$) Suppose that $x \in 2^{\mathbb{N}}$ is not weakly 2-random. Let g be a function witnessing this in the sense of Theorem 2.10.

Fix a rational number $\delta > 0$. We are now constructing the function f by induction on the length of words. Let $f(\square) = 0$. Suppose that we have already defined $f(\sigma)$ for some word σ and we want to define $f(\sigma 0)$ and $f(\sigma 1)$. Let $u(\sigma) = \#\{v : v \preceq \sigma \wedge g(v) = 1\}$. If $u(\sigma) = u(\sigma_0^{|\sigma|-2})$, we let $f(\sigma 0) = f(\sigma 1) = 0$. Otherwise, compute the least n such that

$$\frac{1}{|\sigma| + n}\left(\sum_{i=0}^{|\sigma|-1} f(\sigma_0^i) + n\right) > \delta.$$

Let $f(\sigma w) = 1$ for every $w \in 2^i$ with $i \leq n$. It remains to observe, that if g says 1 on only finitely many prefixes then so does f. Consequently, the average of answers given by f on the prefixes of such sequence converges to 0. On the other hand, if g says 1 on infinitely many prefixes then the average of answers given by f is larger than δ infinitely many times (and so, it does not converge to 0). The rest follows from the properties of g.

($\Rightarrow$) Let f be a computable function such that

$$\limsup_{n\to\infty} \frac{\sum_{i=0}^{n} f(x_0^i)}{n+1} > 0$$

and

$$\mu\left(\lim_{n\to\infty} \frac{\sum_{i=0}^{n} f(X_0^i)}{n+1} = 0\right) = 1.$$

Fix $\delta > 0$ be a rational such that

$$\frac{\sum_{i=0}^{n} f(x_0^i)}{n+1} > \delta$$

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

for infinitely many n. We define a computable function g as follows. For each $w \in 2^{<\mathbb{N}}$ let $g(w) = 1$ if and only if

$$\frac{\sum_{i=0}^{n} f(x_0^i)}{n+1} > \delta.$$

By Theorem 2.10, x is not weakly 2-random. □

Furthermore, a stronger version of Proposition 4.16 follows from the previous considerations.

PROPOSITION 4.21. *There exists a λ-random sequence x and a function $f : 2^{<\mathbb{N}} \to \{0,1\}$ such that*

$$\limsup_{n\to\infty} \frac{\sum_{i=0}^{n} f(x_0^i)}{n+1} > 0$$

and for all $y \neq x$

$$\#\{i : f(y_0^i) = 1\} < \infty.$$

As a consequence, even if an estimator converges to some value in the pointwise fashion almost surely, it may happen that it fails to converge in Cesáro averages on some random point. This is true even under the assumption that the expected value of the estimator is bounded. If the estimator gives finitely many nonzero answers almost surely, then the expected value of the limit of answers is zero. In particular, we have the following.

COROLLARY 4.22. *There exists a computable function $g : 2^{<\mathbb{N}} \to \mathbb{R}^+$ such that $lim_{n\to\infty} g(X_0^n)$ exists almost surely and $\mathbb{E}_\lambda(\sup_i |g(X_0^i)|) < \infty$ and for some random sequence x*

$$\limsup_{i\to\infty} \frac{1}{i} \sum_{k=1}^{i} g(x_0^k) > \mathbb{E}_\lambda \left(\limsup_{k\to\infty} g(X_0^k) \right).$$

§5. One additional remark and a question. Let me end this note with a brief remark about universal inductive schemes. In the introduction, I gave the following motivation for algorithmic randomness. We perform some experiment and we want to know whether a sequence of observations comes from a random process. To this end, we take a computable probability measure, say, produced by a Turing machine with index k. Then we fix a notion of algorithmic randomness and finally, we start saying some nontrivial things about properties that a nice sequence of random outcomes should have. But to be honest, it requires a great deal of knowledge to guess that we should be looking at outputs of k-th Turing machine and not of 17-th machine or at some other possible measure. More often than not, we do not have that kind of knowledge. And from a certain philosophical point of view, it may matter not if something is random as per given probability measure—rather we may want to simply know if it is random at all. If so, then perhaps our true goal is not a notion of randomness with respect to a fixed measure but something more general—randomness with respect to a class of measures. This may be done to some extend as shown by [4]). One way to specialize this into a formal question is as follows.

PROBLEM 5.23. *Is there a natural class $\mathcal{C}$ of measures with a non-trivial learning-theoretic definition of randomness with respect to $\mathcal{C}$? For instance, is there a class of measures $\mathcal{C}$*

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

such that a sequence $x \in 2^{\mathbb{N}}$ *is* 1*-random with respect to some measure from* $\mathcal{C}$ *if and only if there is no computable* $f : 2^{<\mathbb{N}} \to \{0, 1\}$ *such that*

$$\#\{i : f(x_0^i) = 1\} = \infty$$

and for every measure μ *from* $\mathcal{C}$ *and every* $n \in \mathbb{N}$

$$\mu(\#\{i : f(X_0^i) = 1\} \geq n) \leq 2^{-n}?$$

The difference between this and learning-theoretic version of Martin-Löf randomness lies in the measure-theoretic condition, namely, here we ask about recognizing properties that are rare not only from a perspective of one measure but universally, for every measure in the class $\mathcal{C}$. I conjecture that such learning-theoretic characterization is possible for the class of computable stationary ergodic measures. My guess is motivated by the known existence of inductive schemes for this class.

Inductive schemes which presuppose only minimal knowledge about the underlying probability measure, are the holy grails of learning theory, statistics, philosophy of science, etc. For example, various nonparametric schemes for empirical inference that are universal in the class of stationary ergodic processes are known, e.g., Ornstein showed the existence of a universal backward estimator of conditional probability [17] and Algoet studied universal procedures for sequential decisions [1]. In general, these schemes achieve optimal performance almost surely on any measure satisfying some general properties (hence, they are called universal).

Universality (e.g., with respect to some class of computable measures) is a strong property. One could wonder if it is strong enough to guarantee convergence on all relevant random sequences. Learning functions from Propositions 4.16 and 4.21 manifest their unusual behavior on exactly 1-random sequence. The uniform measure λ is continuous, hence we can disturb the convergence on a single sequence without worrying about the behavior on the set of full measure. This is true for every continuous measure μ—anything that happens on a singleton only happens with μ-probability zero. Finally, recall the following theorem.

THEOREM 5.24 (Kautz [13]). *If* μ *is a computable measure and for some* $x \in 2^{\mathbb{N}}$ *we have* $\mu(\{x\}) > 0$ *then* x *is computable.*

Consequently, the behavior of the estimator on a unique λ-random point is irrelevant to the probabilistic properties of the estimator such as universality. In other words, universality with respect to some class of computable measures does not guarantee convergence on every 1-random sequence.

Acknowledgements The author is grateful to Łukasz Dębowski and Dariusz Kalociński for their advice. This work was supported by the National Science Centre Poland grant no. 2018/31/B/HS1/04018.

REFERENCES

[1] Algoet, P. H. (1994). The strong law of large numbers for sequential decisions under uncertainty. *IEEE Transactions on Information Theory*, **40**(3), 609–633.

[2] Bailey, D. H. (1976). Sequential Schemes for Classifying and Predicting Ergodic Processes. PhD Thesis, Stanford University.

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press

[3] Bienvenu, L., Day, A. R., Hoyrup, M., Mezhirov, I., & Shen, A. (2012). A constructive version of Birkhoffs ergodic theorem for Martin-Löf random points. *Information and Computation*, **210**, 21–30.

[4] Bienvenu, L., Gács, P., Hoyrup, M., Rojas, C., & Shen, A. (2011). Algorithmic tests and randomness with respect to a class of measures. *Proceedings of the Steklov Institute of Mathematics*, **274**(1), 34.

[5] Billingsley, P. (2008). *Probability and Measure*. New York: John Wiley & Sons.

[6] Breiman, L. (1957). The individual ergodic theorem of information theory. *The Annals of Mathematical Statistics*, **28**(3), 809–811.

[7] Dębowski, Ł. & Steifer, T. (2020). Universal coding and prediction on Martin-Löf ergodic random points. Preprint, arXiv:2005.03627.

[8] Downey, R. G. & Hirschfeldt, D. R. (2010). *Algorithmic Randomness and Complexity*. Theory and Applications of Computability. New York, NY: Springer New York.

[9] Franklin, J., Greenberg, N., Miller, J., & Ng, K. M. (2012). Martin-Löf random points satisfy Birkhoff's ergodic theorem for effectively closed sets. *Proceedings of the American Mathematical Society*, **140**(10), 3623–3628.

[10] Freer, C., Nies, A., Stephan, F., et al. (2014). Algorithmic aspects of Lipschitz functions. *Computability*, **3**(1), 45–61.

[11] Gold, E. M. (1965). Limiting recursion. *The Journal of Symbolic Logic*, **30**(1), 28–48.

[12] Jockusch Jr., C. G. & Schupp, P. E. (2012). Generic computability, Turing degrees, and asymptotic density. *Journal of the London Mathematical Society*, **85**(2), 472–490.

[13] Kautz, S. M. (1991). Degrees of Random Sets. PhD Thesis, Cornell University.

[14] Lattimore, T., & Hutter, M. (2015). On Martin-Löf (non-) convergence of Solomonoff's universal mixture. *Theoretical Computer Science*, **588**, 2–15.

[15] Martin-Löf, P. (1966). The definition of random sequences. *Information and Control*, **9**(6), 602–619.

[16] Milovanov, A. (2020). Predictions and algorithmic statistics for infinite sequence. Preprint, arXiv:2005.03467.

[17] Ornstein, D. S. (1978). Guessing the next output of a stationary process. *Israel Journal of Mathematics*, **30**(3), 292–296.

[18] Osherson, D. & Weinstein, S. (2008). Recognizing strong random reals. *The Review of Symbolic Logic*, **1**(1), 56–63.

[19] Pathak, N., Rojas, C., & Simpson, S. (2014). Schnorr randomness and the Lebesgue differentiation theorem. *Proceedings of the American Mathematical Society*, **142**(1), 335–349.

[20] Peyerimhoff, A. (2006). *Lectures on Summability*, Vol. 107. Berlin: Springer.

[21] Putnam, H. (1965). Trial and error predicates and the solution to a problem of Mostowski. *The Journal of Symbolic Logic*, **30**(1), 49–57.

[22] Solomonoff, R. J. (1964a). A formal theory of inductive inference. Part I. *Information and Control*, **7**(1), 1–22.

[23] _____ (1964b). A formal theory of inductive inference. Part II. *Information and Control*, **7**(2), 224–254.

[24] Takahashi, H. (2008). On a definition of random sequences with respect to conditional probability. *Information and Computation*, **206**(12), 1375–1382.

[25] V'yugin, V. V. (1998). Ergodic theorems for individual random sequences. *Theoretical Computer Science*, **207**(2), 343–361.

[26] Van Lambalgen, M. (1987). *Random Sequences*. PhD Thesis, University of Amsterdam.

[27] Zaffora Blando, F. (2019). A learning-theoretic characterisation of Martin-Löf randomness and Schnorr randomness. *The Review of Symbolic Logic*, 1–21.

INSTITUTE OF FUNDAMENTAL TECHNOLOGICAL RESEARCH
POLISH ACADEMY OF SCIENCES
UL. PAWINSKIEGO 5B, 02-106, WARSZAWA, POLAND
E-mail: tsteifer@ippt.pan.pl

https://doi.org/10.1017/S1755020321000125 Published online by Cambridge University Press