UNIVERSAL CODING AND PREDICTION
ON ERGODIC RANDOM POINTS

ŁUKASZ DĘBOWSKI [ID] AND TOMASZ STEIFER [ID]

Abstract. Suppose that we have a method which estimates the conditional probabilities of some unknown stochastic source and we use it to guess which of the outcomes will happen. We want to make a correct guess as often as it is possible. What estimators are good for this? In this work, we consider estimators given by a familiar notion of universal coding for stationary ergodic measures, while working in the framework of algorithmic randomness, i.e., we are particularly interested in prediction of Martin-Löf random points. We outline the general theory and exhibit some counterexamples. Completing a result of Ryabko from 2009 we also show that universal probability measure in the sense of universal coding induces a universal predictor in the prequential sense. Surprisingly, this implication holds true provided the universal measure does not ascribe too low conditional probabilities to individual symbols. As an example, we show that the Prediction by Partial Matching (PPM) measure satisfies this requirement with a large reserve.

§1. Introduction. A sequence of outcomes $X_1, X_2, \ldots$ coming from a finite alphabet is drawn in a sequential manner from an unknown stochastic source P. At each moment a finite prefix $X_1^n = (X_1, X_2, \ldots, X_n)$ is available. The forecaster has to predict the next outcome using this information. The task may take one of the two following forms. In the first scenario, the forecaster simply makes a guess about the next outcome. The forecaster's performance is then assessed by comparing the guess with the outcome. This scenario satisfies the weak prequential principle of Dawid [12]. In the second case, we allow the forecaster to be uncertain, namely, we ask them to assign a probability value for each of the outcomes. These values may be interpreted as estimates of the conditional probabilities $P(X_{n+1}|X_1^n)$. Various criteria of success may be chosen here such as the quadratic difference of distributions or the Kullback–Leibler divergence. The key aspect of both problems is that we assume limited knowledge about the true probabilities governing the process that we want to forecast. Thus, an admissible solution should achieve the optimal results for an arbitrary process from some general class. For clarity, term "universal predictor" will be used to denote the solution of guessing the outcome, while the solution of estimation of probabilities will

be referred to as "universal estimator," "universal measure," or "universal code" depending on the exact meaning.

The accumulated literature on universal coding and universal prediction is vast, even when we restrict ourselves to interactions between coding and prediction (see, e.g., [1, 20, 30, 44, 45, 50, 53]). To begin, it is known that for a fixed stochastic source P, the optimal prediction is given by the predictor induced by P, i.e., the informed scheme which predicts the outcome with the largest conditional probability $P(X_{n+1}|X_1^n)$ [1]. In particular, we may expect that a good universal estimator should induce a good universal predictor. That being said, the devil is hidden in the details such as what is meant by a "good" universal code, measure, estimator, or predictor.

In this paper, we will assume that the unknown stochastic source P lies in the class of stationary ergodic measures. Moreover, we are concerned with measures which are universal in the information-theoretic sense of universal coding, i.e., the rate of Kullback–Leibler divergence of the estimate and the true measure P vanishes for any stationary ergodic measure. As for universal predictors, we assume that the rate of correct guesses is equal to the respective rate for the predictor induced by measure P. In this setting, a universal measure need not belong to the class of stationary ergodic measures and can be computable, which makes the problem utterly practical. Our framework should be contrasted with universal prediction à la Solomonoff for left-c.e. semimeasures where the universal semimeasure belongs to the class and is not computable [50]. In general, existence of a universal measure for an arbitrary class of probability measures can be linked to separability of the considered class [48].

Now, we can ask the question whether a universal measure in the above sense of universal coding induces a universal predictor. Curiously, this simple question has not been unambiguously answered in the literature (see [37] for a recent survey), although a host of related propositions were compiled by Suzuki [53] and Ryabko [44, 45] (see also [46]). It was shown by Ryabko [45] (see also [46]) that the expected value of the average absolute difference between the conditional probability for a universal measure and the true value $P(X_{n+1}|X_1^n)$ converges to zero for any stationary ergodic measure P. Ryabko [45] showed also that there exists a universal measure that induces a universal predictor. As we argue in this paper, this result does not solve the general problem.

Completing works [44, 45, 53], in this paper we will show that any universal measure R in the sense of universal coding that additionally satisfies a uniform bound

$$- \log R(X_{n+1}|X_1^n) \le \varepsilon_n \sqrt{n/\ln n}, \quad \lim_{n \to \infty} \varepsilon_n = 0 \qquad (1)$$

induces a universal predictor, indeed. On our way, we will use the Breiman ergodic theorem [7] and the Azuma inequality for martingales with bounded increments [3], which is the source of condition (1). It is left open whether this condition is necessary. Fortunately, condition (1) is satisfied by reasonable universal measures such as the Prediction by Partial Matching (PPM)

measure [10, 44, 47], which we also show in this paper. It may be interesting to exhibit universal measures for which this condition fails. There is a large gap between bound (1) and the respective bound for the PPM measure, which begs for further research.

To add more weight and to make the problem interesting from a computational perspective, we consider this topic in the context of algorithmic randomness and we seek for effective versions of probabilistic statements. Effectivization is meant as the research program of reformulating almost sure statements into respective statements about algorithmically random points, i.e., algorithmically random infinite sequences. Any plausible class of random points is of measure one (see [18]), and the effective versions of theorems substitute phrase "almost surely" with "on all algorithmically random points." Usually, randomness in the Martin-Löf sense is the desired goal [35]. In many cases, the standard proofs are already constructive, and effectivization of some theorems asks for developing new proofs, but sometimes the effective versions are false.

In this paper, we will successfully show that the algorithmic randomness theory is mature enough to make the theory of universal coding and prediction for stationary ergodic sources effective in the Martin-Löf sense. The main keys to this success are: the framework for randomness with respect to uncomputable measures by Reimann and Slaman [42, 43], the effective Birkhoff ergodic theorem [6, 21, 55], an effective version of Breiman's ergodic theorem [7], and an effective Azuma theorem, which follows from the Azuma inequality [3] and the result of Solovay (unpublished; see [18])—which we call here the effective Borel–Cantelli lemma. As a little surprise, there is also a negative result concerning universal forward estimators—Theorem 3.13. Not everything can be made effective.

The organization of the paper is as follows. In Section 2, we discuss preliminaries: notation (Section 2.1), stationary and ergodic measures (Section 2.2), algorithmic randomness (Section 2.3), and some known effectivizations (Section 2.4). Section 3 contains main results concerning: universal coding (Section 3.1), universal prediction (Section 3.2), universal predictors induced by universal backward estimators (Section 3.3) and by universal codes (Section 3.4), as well as the PPM measure (Section 3.5), which constitutes a simple example of a universal code and a universal predictor.

§2. Preliminaries. In this section we familiarize the readers with our notation, we recall the concepts of stationary and ergodic measures, we discuss various sorts of algorithmic randomness, and we recall known facts from the effectivization program.

2.1. Notation. Throughout this paper, we consider the standard measurable space $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ of two-sided infinite sequences over a finite alphabet $\mathbb{X} = \{a_1, \dots, a_D\}$, where $D \geq 2$. (Occasionally, we also apply the space of one-sided infinite sequences $(\mathbb{X}^{\mathbb{N}}, \mathcal{X}^{\mathbb{N}})$.) The points of the space are (infinite) sequences $x = (x_i)_{i \in \mathbb{Z}} \in \mathbb{X}^{\mathbb{Z}}$. We also denote (finite) strings $x_j^k = (x_i)_{j \leq i \leq k}$,

where $x_j^{j-1} = \lambda$ equals the empty string. By $\mathbb{X}^* = \bigcup_{n \geq 0} \mathbb{X}^n$ we denote the set of strings of an arbitrary length including the singleton $\mathbb{X}^0 = \{\lambda\}$. We use random variables $X_k((x_i)_{i \in \mathbb{Z}}) := x_k$. Having these, the σ-field $\mathcal{X}^{\mathbb{Z}}$ is generated by cylinder sets $(X_{-|\sigma|+1}^{|\tau|} = \sigma\tau)$ for all $\sigma, \tau \in \mathbb{X}^*$. We tacitly assume that P and R denote probability measures on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$. For any probability measure P, we use the shorthand notations $P(x_1^n) := P(X_1^n = x_1^n)$ and $P(x_j^n | x_1^{j-1}) := P(X_j^n = x_j^n | X_1^{j-1} = x_1^{j-1})$. Notation $\log x$ denotes the binary logarithm, whereas $\ln x$ is the natural logarithm.

2.2. Stationary and ergodic measures. Let us denote the measurable shift operation $T((x_i)_{i \in \mathbb{Z}}) := (x_{i+1})_{i \in \mathbb{Z}}$ for two-sided infinite sequences $(x_i)_{i \in \mathbb{Z}} \in \mathbb{X}^{\mathbb{Z}}$.

DEFINITION 2.1 (Stationary measures). A probability measure P on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ is called stationary if $P(T^{-1}(A)) = P(A)$ for all events $A \in \mathcal{X}^{\mathbb{Z}}$.

DEFINITION 2.2 (Ergodic measures). A probability measure P on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$ is called ergodic if for each event $A \in \mathcal{X}^{\mathbb{Z}}$ such that $T^{-1}(A) = A$ we have either $P(A) = 1$ or $P(A) = 0$.

The class of stationary ergodic probability measures has various nice properties guaranteed by the collection of fundamental results called ergodic theorems. Typically stationary ergodic measures are not computable (e.g., consider independent biased coin tosses with a common uncomputable bias) but they allow for computable universal coding and computable universal prediction schemes can achieve optimal error rates, as it will be explained in Section 3.

2.3. Sorts of randomness. Now let us discuss some computability notions. In the following, *computably enumerable* is abbreviated as *c.e.* Given a real r, the set $\{q \in \mathbb{Q} : q < r\}$ is called the left cut of r. A real function f with arguments in a countable set is called computable or left-c.e. respectively if the left cuts of $f(\sigma)$ are uniformly computable or c.e. given an enumeration of σ. For an infinite sequence $s \in \mathbb{X}^{\mathbb{Z}}$, we say that real functions f are *s*-computable or *s*-left-c.e. if they are computable or left-c.e. with oracle s. Similarly, for a real function f taking arguments in $\mathbb{X}^{\mathbb{Z}}$, we will say that f is *s*-computable or *s*-left-c.e. if left cuts of $f(x)$ are uniformly computable or c.e. with oracles $x \oplus s := (\ldots, x_{-1}, s_{-1}, x_0, s_0, x_1, s_1, \ldots)$. This induces in effect *s*-computable and *s*-left-c.e. random variables and stochastic processes on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$, where the values of an *s*-computable (*s*-left-c.e.) variable on a point x are $(x \oplus s)$-computable (*s*-left-c.e.) uniformly in x and the values of *s*-computable (*s*-left-c.e.) process X_i (with natural or integer i) are *s*-computable (*s*-left-c.e.).

For stationary ergodic measures, we need a definition of algorithmically random points with respect to an arbitrary, i.e., not necessarily computable probability measure on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$. A simple definition thereof was proposed by Reimann [42] and Reimann and Slaman [43]. This definition is equivalent to earlier approaches by Levin [32–34] and Gács [23] as shown by Day and Miller [13] and we will use it since it leads to

straightforward generalizations of the results in Section 2.4. The definition is based on measure representations. Let $\mathcal{P}(\mathbb{X}^{\mathbb{Z}})$ be the space of probability measures on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$. A measure $P \in \mathcal{P}(\mathbb{X}^{\mathbb{Z}})$ is called s-computable if real function $(\sigma, \tau) \mapsto P(X_{-|\sigma|+1}^{|\tau|} = \sigma\tau)$ is s-computable. Similarly, a representation function is a function $\rho : \mathbb{X}^{\mathbb{Z}} \to \mathcal{P}(\mathbb{X}^{\mathbb{Z}})$ such that real function $(\sigma, \tau, s) \mapsto \rho(s)(X_{-|\sigma|+1}^{|\tau|} = \sigma\tau)$ is computable. Subsequently, we say that an infinite sequence $s \in \mathbb{X}^{\mathbb{Z}}$ is a representation of measure P if there exists a representation function ρ such that $\rho(s) = P$. We note that any measure P is s-computable for any representation s of P.

We will consider two important sorts of algorithmically random points: Martin-Löf or 1-random points and weakly 2-random points with respect to an arbitrary stationary ergodic measure P on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{X}^{\mathbb{Z}})$. Note that the following notions are typically defined for one-sided infinite sequences over the binary alphabet and computable measures P. In the following parts of this paper, let an infinite sequence $s \in \mathbb{X}^{\mathbb{Z}}$ be a representation of measure P.

DEFINITION 2.3. A collection of events $U_1, U_2, \dots \in \mathcal{X}^{\mathbb{Z}}$ is called uniformly s-c.e. if and only if there is a collection of sets $V_1, V_2, \dots \subset \mathbb{X}^* \times \mathbb{X}^*$ such that

$$U_i = \left\{ x \in \mathbb{X}^{\mathbb{Z}} : \exists (\sigma, \tau) \in V_i : x_{-|\sigma|+1}^{|\tau|} = \sigma\tau \right\}$$

and sets $V_1, V_2, \dots$ are uniformly s-c.e.

DEFINITION 2.4 (Martin-Löf test). A uniformly s-c.e. collection of events $U_1, U_2, \dots \in \mathcal{X}^{\mathbb{Z}}$ is called a Martin-Löf (s, P)-test if $P(U_n) \leq 2^{-n}$ for every $n \in \mathbb{N}$.

DEFINITION 2.5 (Martin-Löf or 1-randomness). A point $x \in \mathbb{X}^{\mathbb{Z}}$ is called Martin-Löf (s, P)-random or 1-(s, P)-random if for each Martin-Löf (s, P)-test $U_1, U_2, \dots$ we have $x \notin \bigcap_{i \geq 1} U_i$. A point is called Martin-Löf P-random or 1-P-random if it is 1-(s, P)-random for some representation s of P.

Subsequently, an event $C \in \mathcal{X}^{\mathbb{Z}}$ is called a $\Sigma_2^0(s)$ event if there exists a uniformly s-c.e. sequence of events $U_1, U_2, \dots$ such that $\mathbb{X}^{\mathbb{Z}} \setminus C = \bigcap_{i \geq 1} U_i$.

DEFINITION 2.6 (Weak 2-randomness). A point $x \in \mathbb{X}^{\mathbb{Z}}$ is called weakly 2-(s, P)-random if x is contained in every $\Sigma_2^0(s)$ event C such that $P(C) = 1$. A point is called weakly 2-P-random if it is weakly 2-(s, P)-random for some representation s of P.

The sets of weakly 2-random points are strictly smaller than the respective sets of 1-random points (see [18]).

In general, there is a whole hierarchy of algorithmically random points, such as (weakly) n-random points, where n runs over natural numbers. For our purposes, however, only 1-random points and weakly 2-random points matter since the following proposition sets the baseline for effectivization:

PROPOSITION 2.7 (Folklore). *Let $Y_1, Y_2, \dots$ be a sequence of uniformly s-computable random variables. If limit $\lim_{n \to \infty} Y_n$ exists P-almost surely, then it exists on all weakly 2-(s, P)-random points.*

The above proposition is obvious since the set of points on which limit $\lim_{n\to\infty} Y_n$ exists is a $\Sigma_2^0(s)$ event. The effectivization program aims to strengthen the above claim to 1-P-random points (or even weaker notions such as Schnorr randomness) but this need not always be feasible. In particular, one can observe that:

PROPOSITION 2.8 (Folklore). *Let P be a non-atomic computable measure on $\mathbb{X}^{\mathbb{N}}$. Then there exists a computable function $f : \mathbb{X}^* \to \{0, 1\}$ such that the limit $\lim_{n\to\infty} f(X_1^n)$ exists and is equal zero P-almost surely but it is not defined on exactly one point, which is 1-P-random.*

This fact is a simple consequence of the existence of Δ_2^0 1-P-random sequences (for a computable P) and may be also interpreted in terms of learning theory (cf. [41] and the upcoming paper [52]).

2.4. Known effectivizations. Many probabilistic theorems have been effectivized so far. Usually they were stated for computable measures but their generalizations for uncomputable measures follow easily by relativization, i.e., putting a representation s of measure P into the oracle. In this section, we list several known effectivizations of almost sure theorems which we will use further.

As shown by Solovay (unpublished; see [18]), we have this effective version of the Borel–Cantelli lemma:

PROPOSITION 2.9 (Effective Borel–Cantelli lemma). *Let P be a probability measure. If a uniformly s-c.e. sequence of events $U_0, U_1, \ldots \in \mathcal{X}^{\mathbb{Z}}$ satisfies $\sum_{i=1}^{\infty} P(U_n) < \infty$ then $\sum_{i=1}^{\infty} \mathbf{1}\{x \in U_n\} < \infty$ on each 1-(s, P)-random point x.*

By the effective Borel–Cantelli lemma, Proposition 2.9 follows the effective version of the Barron lemma [5, Theorem 3.1]:

PROPOSITION 2.10 (Effective Barron lemma). *For any probability measure P and any s-computable probability measure R, on 1-(s, P)-random points we have*

$$\lim_{n\to\infty} \left[-\log R(X_1^n) + \log P(X_1^n) + 2\log n \right] = \infty. \qquad (2)$$

In the following, we make an easy but important observation—probabilities conditioned on an infinite past are defined on random points. First, we need to recall the notion of a martingale process and prove an effective version of Doob's martingale convergence.

DEFINITION 2.11 (Martingale process). A process $(X_i)_{i\in\mathbb{N}}$ is called a martingale process relative to the sequence of σ-algebras $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ (called a filtration) if the following conditions hold:

1. X_n are $\mathcal{F}_n$-measurable for all n;
2. $\mathbf{E}(|X_n|) < \infty$ for all n;
3. $\mathbf{E}(X_{n+1}|\mathcal{F}_n) = X_n$ for all n almost surely.

The proof of Doob's martingale convergence can be easily made effective. This was already observed by Takahashi [54], who stated the effective

martingale convergence for a specific filtration generated by cylinders X_1^n. The following upcrossing inequality can be used to define a test which enforces convergence.

PROPOSITION 2.12 (Doob upcrossing inequality). *Let $(X_i)_{i\in\mathbb{N}}$ be a martingale process and let C_n will be the random variable denoting the number of upcrossings of interval $[a, b]$ (with $a, b \in \mathbb{R}$) by time n and suppose that $\sup_n \boldsymbol{E}(|X_n|) < \infty$. Then for each n, we have*

$$\boldsymbol{E}\left(\sup_n C_n\right) \le \frac{|a| + \sup_n \boldsymbol{E}(|X_n|)}{b - a}.$$

PROPOSITION 2.13 (Effective Doob martingale convergence). *Let $(X_i)_{i\in\mathbb{N}}$ be a uniformly s-computable martingale process with $\sup_n \boldsymbol{E}(|X_n|) < \infty$. Then, limit $\lim_{n\to\infty} X_n$ exists and is finite on each 1-(s, P)-random point.*

PROOF. Suppose that process $(X_i)_{i\in\mathbb{N}}$ does not converge on some random point x. Then there exist rational a, b such that the number of upcrossings of the interval $[a, b]$ by $X_i(x)$ is infinite. Let C_n be the random variable denoting the number of upcrossings of interval $[a, b]$ by the process $(X_i)_{i\in\mathbb{N}}$ by the time n. Let C_∞ denote $\sup_n C_n$ and let $f : \mathbb{N} \to \mathbb{N}$ be a monotonic function. Consider a collection of sets $U_1, U_2, \dots$ such that for all $i > 0$

$$U_i = \{\omega : C_\infty(\omega) > f(n)\}.$$

By Proposition 2.12 (Doob upcrossing inequality) and the Markov inequality, we have

$$P(U_i) \le \frac{|a| + \sup_n \mathbf{E}(|X_n|)}{f(n)(b - a)}.$$

Note that if f grows sufficiently fast, then $\sum_{i=1}^{\infty} P(U_i)$ converges. Moreover, the collection of sets $U_1, U_2, \dots$ is uniformly s-c.e. It follows by Proposition 2.9 (effective Borel–Cantelli lemma) that $C_\infty(x) < \infty$ for every 1-(s, P)-random point x, which is a contradiction.

It remains to observe that the limit of $(X_i)_{i\in\mathbb{N}}$ is finite. This follows easily if one considers the collection of sets $V_1, V_2, \dots$ with

$$V_i = \{x : \sup_n X_n(x) > 2^n\},$$

which are uniformly s-c.e. By the Markov inequality and the monotone convergence theorem, we have $P(V_i) \le 2^{-i} \sup_n \mathbf{E}(X_n)$. We apply Proposition 2.9 to conclude that X_n are bounded on every 1-(s, P)-random point. $\dashv$

Random variables $P(x_0|X_{-n}^{-1})$ for $n \ge 1$ form a uniformly s-computable martingale process with respect to the filtration generated by cylinder sets X_{-n}^{-1} for any representation s of P. Thus applying the effective Doob martingale convergence, we obtain an effective version of the Lévy law in particular. In this work, our attention is limited to the following form.

PROPOSITION 2.14 (Effective Lévy law). *On 1-P-random points there exist limits*

$$P(x_0|X_{-\infty}^{-1}) := \lim_{n\to\infty} P(x_0|X_{-n}^{-1}). \tag{3}$$

Now let us proceed to a celebrated result of the algorithmic randomness theory, which is the effective Birkhoff ergodic theorem [6, 21, 28, 29, 39, 55]. In the following, $\mathbf{E}\,X := \int X dP$ stands for the expectation of a random variable X with respect to measure P.

PROPOSITION 2.15 (Effective Birkhoff ergodic theorem [6, Theorem 10]). *For a stationary ergodic probability measure P and an s-left-c.e. real random variable G such that $G \geq 0$ and $\mathbf{E}\,G < \infty$, on 1-(s, P)-random points we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} G \circ T^i = \mathbf{E}\,G. \tag{4}$$

We note in passing that if a point is not 1-random for a computable P then (4) fails on this point for some computable real random variable G and some computable transformation T [22].

The proof of the next proposition is an easy application of Proposition 2.15 and properties of left-c.e. functions.

PROPOSITION 2.16 (Effective Breiman ergodic theorem [51]). *For a stationary ergodic probability measure P and uniformly s-computable real random variables $(G_i)_{i\geq 0}$ such that $G_n \geq 0$, $\mathbf{E}\sup_n G_n < \infty$, and limit $\lim_{n\to\infty} G_n$ exists P-almost surely, on 1-(s, P)-random points we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} G_i \circ T^i = \mathbf{E} \lim_{n\to\infty} G_n. \tag{5}$$

PROOF. Let $H_k := \sup_{t>k} G_t \geq 0$. Then $G_t \leq H_k$ for all $t > k$ and consequently,

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} G_i \circ T^i \leq \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} H_k \circ T^i. \tag{6}$$

Observe that the supremum H_k of uniformly s-computable functions $G_{k+1}, G_{k+2}, \ldots$ is s-left-c.e. Indeed, to enumerate the left cut of the supremum $H_k(x)$ we simultaneously enumerate the left cuts of $G_{k+1}(x), G_{k+2}(x), \ldots$. This is possible since every s-computable function is also s-left-c.e. Moreover, we are considering only countably many functions, and hence we can guarantee that an element of each left cut appears in the enumeration at least once.

Now, since for all $k \geq 0$ random variables H_k are s-left-c.e. then by Theorem 2.15 (effective Birkhoff ergodic theorem), on 1-(s, P)-random points we have

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} G_i \circ T^i \leq \mathbf{E}\,H_k. \tag{7}$$

Since $H_k \geq 0$ and $\mathbf{E} \sup_k H_k < \infty$ then by the dominated convergence,

$$\inf_{k \geq 0} \mathbf{E} H_k = \lim_{k \to \infty} \mathbf{E} H_k = \mathbf{E} \lim_{k \to \infty} H_k = \mathbf{E} \lim_{n \to \infty} G_n. \tag{8}$$

Thus,

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} G_i \circ T^i \leq \mathbf{E} \lim_{n \to \infty} G_n. \tag{9}$$

For the converse inequality, consider a natural number M and put random variables $\bar{H}_k := M - \inf_{t > k} \min \{G_t, M\} \in [0, M]$. We observe that $\bar{H}_k$ are also s-left-c.e. since G_t are uniformly s-computable by the hypothesis. By Theorem 2.15 (effective Birkhoff ergodic theorem), on 1-(s, P)-random points we have

$$M - \liminf_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \min \{G_i, M\} \circ T^i \leq \mathbf{E} \bar{H}_k. \tag{10}$$

Since $0 \leq \bar{H}_k \leq M$ then by the dominated convergence,

$$\inf_{k \geq 0} \mathbf{E} \bar{H}_k = \lim_{k \to \infty} \mathbf{E} \bar{H}_k = \mathbf{E} \lim_{k \to \infty} \bar{H}_k = M - \mathbf{E} \min \left\{ \lim_{n \to \infty} G_n, M \right\}. \tag{11}$$

Hence, regrouping the terms we obtain

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} G_i \circ T^i \geq \liminf_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \min \{G_i, M\} \circ T^i$$

$$\geq \mathbf{E} \min \left\{ \lim_{n \to \infty} G_n, M \right\} \xrightarrow[M \to \infty]{} \mathbf{E} \lim_{n \to \infty} G_n, \tag{12}$$

where the last transition follows by the monotone convergence. By (9) and (12) we derive the claim. ⊣

The almost sure versions of Propositions 2.15 and 2.16 concern random variables which need not be nonnegative [7].

An important result for universal prediction is the Azuma inequality [3], whose following corollary will be used in Sections 3.2 and 3.4.

THEOREM 2.17 (Effective Azuma theorem). *For a probability measure P and uniformly s-computable real random variables $(Z_n)_{n \geq 1}$ such that $Z_n = g(X_1^n, s)$ and $|Z_n| \leq \varepsilon_n \sqrt{n / \ln n}$ with $\lim_{n \to \infty} \varepsilon_n = 0$, on 1-$(s, P)$-random points we have*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \left[Z_i - \mathbf{E} \left(Z_i | X_1^{i-1} \right) \right] = 0. \tag{13}$$

PROOF. Define

$$Y_n := \sum_{i=1}^{n} \left[Z_i - \mathbf{E} \left(Z_i | X_1^{i-1} \right) \right]. \tag{14}$$

Process $(Y_n)_{n\geq 1}$ is a martingale with respect to process $(X_n)_{n\geq 1}$ with increments bounded by inequality

$$\left| Z_n - \mathbf{E}\left(Z_n \middle| X_1^{n-1}\right)\right| \leq c_n := 2\varepsilon_n \sqrt{n/\ln n}. \qquad (15)$$

By the Azuma inequality [3] for any $\varepsilon > 0$ we obtain

$$P(|Y_n| \geq n\varepsilon) \leq 2\exp\left(-\frac{\varepsilon^2 n^2}{2\sum_{i=1}^n c_i^2}\right) \leq n^{-\alpha_n}, \qquad (16)$$

where

$$\alpha_n := \frac{\varepsilon^2 n}{8\sum_{i=1}^n \varepsilon_i^2}. \qquad (17)$$

Since $\alpha_n \to \infty$, we have $\sum_{n=1}^{\infty} P(|Y_n| \geq n\varepsilon) < \infty$ and by Proposition 2.9 (effective Borel–Cantelli lemma), we obtain (13) on 1-(s, P)-random points. ⊣

§3. Main results. This section contains results concerning effective universal coding and prediction, predictors induced by universal measures, and some examples of universal measures and universal predictors.

3.1. Universal coding. Let us begin our considerations with the problem of universal measures, which is related to the problem of universal coding. Suppose that we want to compress losslessly a typical sequence generated by a stationary probability measure P. We can reasonably ask what is the lower limit of such a compression, i.e., what is the minimal ratio of the encoded string length divided by the original string length. In information theory, it is well known that the greatest lower bound of such ratios is given by the entropy rate of measure P. For a stationary probability measure P, we denote its entropy rate as

$$h_P := \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left[-\log P(X_1^n)\right] = \lim_{k\to\infty}\mathbf{E}\left[-\log P(X_{k+1}|X_1^k)\right], \qquad (18)$$

which exists for any stationary probability measure.

The entropy rate has the interpretation of the minimal asymptotic rate of lossless encoding of sequences emitted by measure P in various senses: in expectation, almost surely, or on algorithmically random points, where the last interpretation will be pursued in this subsection.

To furnish some theoretical background for universal coding let us recall the Kraft inequality $\sum_{w\in A} 2^{-|w|} \leq 1$, which holds for any prefix-free subset of strings $A \subset \{0, 1\}^*$. The Kraft inequality implies in particular that lossless compression procedures, called prefix-free codes, can be mapped one-to-one to semi-measures. In particular, if we are seeking for a universal code, i.e., a prefix-free code $w \mapsto C(w) \in \{0, 1\}^*$ which is optimal for some class of stochastic sources P, we can equivalently seek for a universal semi-measure of form $w \mapsto R(w) := 2^{-|C(w)|}$. (Similar correspondence holds also for uniquely decodable codes [36].) Consequently, the problem of universal

coding will be solved if we point out such a semi-measure R that

$$\lim_{n\to\infty} \frac{1}{n} \left[-\log R(X_1^n) \right] = \lim_{n\to\infty} \frac{1}{n} \left| C(X_1^n) \right| = h_P \qquad (19)$$

for some points that are typical of P.

As it is well established in information theory, some initial insight into the problem of universal coding or universal measures is given by the Shannon–McMillan–Breiman (SMB) theorem, which states that function $\frac{1}{n} \left[-\log P(X_1^n) \right]$ tends P-almost surely to the entropy rate h_P. The classical proofs of this result were given by Algoet and Cover [2] and Chung [9]. An effective version of the SMB theorem was presented by Hochman [26] and Hoyrup [27] (cf. [38, 55] for related partial and weaker results).

THEOREM 3.1 (Effective SMB theorem [26, 27]). *For a stationary ergodic probability measure* P, *on* 1-P-*random points we have*

$$\lim_{n\to\infty} \frac{1}{n} \left[-\log P(X_1^n) \right] = h_P. \qquad (20)$$

The essential idea of Hoyrup's proof, which is a bit more complicated, can be retold using tools developed in Section 2.4. Observe first that we have

$$\frac{1}{n} \left[-\log P(X_1^n) \right] = \frac{1}{n} \sum_{i=1}^{n} \left[-\log P(X_i | X_1^{i-1} \cdot) \right]. \qquad (21)$$

Moreover, we have the uniform bound

$$\mathbf{E} \sup_{n\geq 0} \left[-\log P(X_0 | X_{-n}^{-1}) \right] \leq \mathbf{E} \left[-\log P(X_0) \right] + \log e \leq \log eD < \infty \qquad (22)$$

(see [49, Lemma 4.26])—invoked by Hoyrup as well. Consequently, the effective SMB theorem follows by Proposition 2.16 (effective Breiman ergodic theorem) and Proposition 2.14 (effective Lévy law). In contrast, the reasoning by Hoyrup was more casuistic and his effective version of the Breiman ergodic theorem is weaker than the one proven here.

We note in passing that it could be also interesting to check whether one can effectivize the textbook sandwich proof of the SMB theorem by Algoet and Cover [2] using the decomposition of conditionally algorithmically random sequences by Takahashi [54]. However, this step would require some novel theoretical considerations about conditional algorithmic randomness for uncomputable measures. We mention this only to point out a possible direction for future research.

As a direct consequence of the effective SMB theorem and Proposition 2.10 (effective Barron lemma), we obtain this effectivization of another well-known almost sure statement:

THEOREM 3.2 (Effective source coding). *For any stationary ergodic measure* P *and any* s-*computable probability measure* R, *on* 1-(s, P)-*random points we have*

$$\liminf_{n\to\infty} \frac{1}{n} \left[-\log R(X_1^n) \right] \geq h_P. \qquad (23)$$

In the almost sure setting, relationship (23) holds *P*-almost surely for any stationary ergodic measure *P* and any (not necessarily computable) probability measure *R*.

Now we can define universal measures.

DEFINITION 3.3 (Universal measure). A computable (not necessarily stationary) probability measure *R* is called (weakly) *n*-universal if for any stationary ergodic probability measure *P*, on (weakly) *n*-*P*-random points we have

$$\lim_{n \to \infty} \frac{1}{n} \left[-\log R(X_1^n) \right] = h_P. \tag{24}$$

In the almost sure setting, we say that a probability measure *R* is almost surely universal if (24) holds *P*-almost surely for any stationary ergodic probability measure *P*. By Proposition 2.7, there are only two practically interesting cases of computable universal measures: weakly 2-universal ones and 1-universal ones, since every computable almost surely universal probability measure is automatically weakly 2-universal. We stress that we impose computability of (weakly) *n*-universal measures by definition since it simplifies statements of some theorems. This should be contrasted with universal prediction à la Solomonoff for left-c.e. semimeasures where the universal element belongs to the class and is not computable [50].

Computable almost surely universal measures exist if the alphabet $\mathbb{X}$ is finite. An important example of an almost surely universal and, as we will see in Section 3.5, also 1-universal measure is the Prediction by Partial Matching (PPM) measure [10, 44, 47]. As we have mentioned, universal measures are closely related to the problem of universal coding (data compression) and more examples of universal measures can be constructed from universal codes, for instance given in [8, 14, 31, 56], using the normalization by Ryabko [45]. This normalization is not completely straightforward, since we need to forge semi-measures into probability measures.

3.2. Universal prediction. Universal prediction is a problem similar to universal coding. In this problem, we also seek for a single procedure that would be optimal within a class of probabilistic sources but we apply a different loss function, namely, we impose the error rate being the density of incorrect guesses of the next output given previous ones. In spite of this difference, we will try to state the problem of universal prediction analogously to universal coding. A predictor is an arbitrary total function $f : \mathbb{X}^* \to \mathbb{X}$. The predictor induced by a probability measure *P* will be defined as

$$f_P(x_1^n) := \arg\max_{x_{n+1} \in \mathbb{X}} P(x_{n+1}|x_1^n), \tag{25}$$

where $\arg\max_{x \in \mathbb{X}} g(x) := \min \left\{ a \in \mathbb{X} : g(a) \geq g(x) \text{ for all } x \in \mathbb{X} \right\}$ for the total order $a_1 < \cdots < a_D$ on $\mathbb{X} = \{a_1, \ldots, a_D\}$. Moreover, for a stationary

measure P, we define the unpredictability rate

$$u_P := \lim_{n \to \infty} \mathbf{E}\left[1 - \max_{x_0 \in \mathbb{X}} P(x_0|X_{-n}^{-1})\right]. \tag{26}$$

It is natural to ask whether the unpredictability rate can be related to entropy rate. Using the Fano inequality [19], a classical result of information theory, and its converse [16], both independently brought to computability theory by Fortnow and Lutz [20], yields this bound:

THEOREM 3.4. *For a stationary measure P over a D-element alphabet,*

$$\frac{D}{D-1}\eta\left(\frac{1}{D}\right)u_P \le h_P \le \eta(u_P) + u_P \log(D-1), \tag{27}$$

where $\eta(p) := - p \log p - (1-p) \log(1-p)$.

Moreover, Fortnow and Lutz [20] found out some stronger inequalities, sandwich-bounding the unpredictability of an arbitrary sequence in terms of its effective dimension. The effective dimension turns out to be a generalization of the entropy rate to arbitrary sequences [27], which are not necessarily random with respect to stationary ergodic measures.

In the less general framework of stationary ergodic measures, using the Azuma theorem, we can show that no predictor can beat the induced predictor and the error rate committed by the latter equals the unpredictability rate u_P. The following proposition concerning the error rates effectivizes the well-known almost sure proposition (the proof in the almost sure setting is available in [1]).

THEOREM 3.5 (Effective source prediction). *For any stationary ergodic measure P and any s-computable predictor f, on* 1-(s, P)-*random points we have*

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}\{X_{i+1} \ne f(X_1^i)\} \ge u_P. \tag{28}$$

Moreover, if the induced predictor f_P is s-computable then (28) *holds with the equality for* $f = f_P$.

PROOF. Let measure P be stationary ergodic. In view of Theorem 2.17 (effective Azuma theorem), for any s-computable predictor f, on 1-(s, P)-random points we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[\mathbf{1}\{X_{i+1} \ne f(X_1^i)\} - P(X_{i+1} \ne f(X_1^i)|X_1^i)\right] = 0. \tag{29}$$

Moreover, we have

$$P(X_{i+1} \ne f(X_1^i)|X_1^i) \ge 1 - \max_{x_{i+1} \in \mathbb{X}} P(x_{i+1}|X_1^i). \tag{30}$$

Subsequently, we observe that limits $\lim_{n \to \infty} P(x_0|X_{-n}^{-1})$ exist on 1-(s, P)-random points by Proposition 2.14 (effective Lévy law). Thus by

Proposition 2.16 (effective Breiman ergodic theorem) and the dominated convergence, on 1-(s, P)-random points we obtain

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[1 - \max_{x_{i+1}\in\mathbb{X}} P(x_{i+1}|X_1^i) \right] = \mathbf{E} \lim_{n\to\infty} \left[1 - \max_{x_0\in\mathbb{X}} P(x_0|X_{-n}^{-1}) \right]$$
$$= u_P. \tag{31}$$

Hence inequality (28) follows by (29)–(31). Similarly, the equality in (28) for $f = f_P$ follows by noticing that inequality (30) turns out to be the equality in this case. ⊣

In the almost sure setting, relationship (28) holds P-almost surely for any stationary ergodic measure P and any (not necessarily computable) predictor f.

We can see that there can be some problem in the effectivization of relationship (28) caused by the induced predictor f_P possibly not being s-computable for certain representations s of measure P—since sometimes testing the equality of two real numbers cannot be done in a finite time. However, probabilities $P(X_{i+1} \neq f_P(X_1^i)|X_1^i)$ are always s-computable. Thus, we can try to define universal predictors in the following way.

DEFINITION 3.6 (Universal predictor). A computable predictor f is called (weakly) n-universal if for any stationary ergodic probability measure P, on (weakly) n-P-random points we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}\{X_{i+1} \neq f(X_1^i)\} = u_P. \tag{32}$$

In the almost sure setting, we say that a predictor f is almost surely universal if (32) holds P-almost surely for any stationary ergodic probability measure P. Almost surely universal predictors exist if the alphabet $\mathbb{X}$ is finite [1, 4, 24, 25, 40]. In [51] it was proved that the almost sure predictor by [25] is also 1-universal.

3.3. Predictors induced by backward estimators. The almost surely universal predictors by [1, 4, 24, 25, 40] were constructed without a reference to universal measures. Nevertheless, these constructions are all based on estimation of conditional probabilities. For a stationary ergodic process one can consider two separate problems: backward and forward estimation. The first problem is naturally connected to prediction. We want to estimate the conditional probability of $(n + 1)$-th bit given the first n bits. Is it possible that, as we increase n, our estimates converge to the true value at some point? To be precise, we ask whether there exists a probability measure R such that for every stationary ergodic measure P we have P-almost surely

$$\lim_{n\to\infty} \sum_{x_{n+1}\in\mathbb{X}} \left| R(x_{n+1}|X_1^{n-1}) - P(x_{n+1}|X_1^{n-1}) \right| = 0. \tag{33}$$

It was shown by Bailey [4] that this is not possible. As we are about to see, we can get something a bit weaker, namely, the convergence in Cesaro averages. But to get there, it will be helpful to consider a bit different problem.

Suppose again that we want to estimate a conditional probability but the bit that we are interested in is fixed and we are looking more and more into the past. In this scenario, we want to estimate the conditional probability $P(x_0|X_{-\infty}^{-1})$ and we ask whether increasing the knowledge of the past can help us achieve the perfect guess. Precisely, we ask if there exists a probability measure R such that for every stationary ergodic measure P we have P-almost surely

$$\lim_{n\to\infty} \sum_{x_0\in\mathbb{X}} \left| R(x_0|X_{-n}^{-1}) - P(x_0|X_{-\infty}^{-1}) \right| = 0. \tag{34}$$

It was famously shown by Ornstein that such estimators exist. (Ornstein proved this for binary-valued processes but the technique can be generalized to finite-valued processes.)

THEOREM 3.7 (Ornstein theorem [40]). *Let the alphabet be finite. There exists a computable measure R such that for every stationary ergodic measure P we have P-almost surely that*

$$\lim_{n\to\infty} \sum_{x_0\in\mathbb{X}} \left| R(x_0|X_{-n}^{-1}) - P(x_0|X_{-\infty}^{-1}) \right| = 0. \tag{35}$$

DEFINITION 3.8. We call a measure R an almost surely universal backward estimator when it satisfies condition (35) P-almost surely for every stationary ergodic measure P, whereas it is called a (weakly) n-universal backward estimator if R is computable and convergence (35) holds on all respective (weakly) n-P-random points.

One can come up with a naive idea: What if we take a universal backward estimator and use it in a forward fashion? Surprisingly, this simple trick gives us almost everything we can get, i.e., a forward estimator that converges to the conditional probability on average. Bailey [4] showed that for an almost surely universal backward estimator R and for every stationary ergodic measure P we have P-almost surely

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \sum_{x_{i+1}\in\mathbb{X}} \left| R(x_{i+1}|X_1^i) - P(x_{i+1}|X_1^i) \right| = 0. \tag{36}$$

The proof of this fact is a direct application of the Breiman ergodic theorem. Since we have a stronger effective version of the Breiman theorem (Theorem 2.16), we can strengthen Bailey's result to an effective version as well. It turns out that even if we take a backward estimator that is good only almost surely (possibly failing on some random points), then the respective result for the forward estimation will hold in the strong sense—on every 1-P-random point.

THEOREM 3.9 (Effective Bailey theorem). *Let R be a computable almost surely universal backward estimator. For every stationary ergodic measure P on 1-P-random points we have* (36).

PROOF. Let R be a computable almost surely universal backward estimator. Fix an $x \in \mathbb{X}$. By Proposition 2.14 (effective Lévy law), for every stationary ergodic probability measure P we have P-almost surely

$$\lim_{n \to \infty} \left| R(x|X_{-n}^{-1}) - P(x|X_{-n}^{-1}) \right| = 0. \tag{37}$$

Note that the bound $0 \le \left| R(x|X_{-n}^{-1}) - P(x|X_{-n}^{-1}) \right| \le 1$ holds uniformly. Moreover, variables $R(x|X_{-n}^{-1}) - P(x|X_{-n}^{-1})$ are uniformly s-computable for any representation s of P. Hence, we can apply Theorem 2.16 (effective Breiman ergodic theorem) to obtain

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \left| R(x|X_1^i) - P(x|X_1^i) \right| = \mathbf{E}\, 0 = 0 \tag{38}$$

for 1-P-random points. The claim follows from this immediately. ⊣

DEFINITION 3.10. We call a measure R an almost surely universal forward estimator when it satisfies condition (36) P-almost surely for every stationary ergodic measure P, whereas it is called a (weakly) n-universal forward estimator if R is computable and convergence (36) holds on all respective (weakly) n-P-random points.

One can expect that the predictor f_R induced by a universal forward estimator R in the sense of Definition 3.10 is also universal in the sense of Definition 3.6. This is indeed true. To show this fact, we will first prove a certain inequality for induced predictors, which generalizes the result from [17, Theorem 2.2] for binary classifiers. This particular observation seems to be new.

PROPOSITION 3.11 (Prediction inequality). *Let p and q be two probability distributions over a countable alphabet $\mathbb{X}$. For $x_p = \arg\max_{x \in \mathbb{X}} p(x)$ and $x_q = \arg\max_{x \in \mathbb{X}} q(x)$, we have inequality*

$$0 \le p(x_p) - p(x_q) \le \sum_{x \in \mathbb{X}} \left| p(x) - q(x) \right|. \tag{39}$$

PROOF. Without loss of generality, assume $x_p \ne x_q$. By the definition of x_p and x_q, we have $p(x_p) - p(x_q) \ge 0$ and $q(x_q) - q(x_p) \ge 0$. Hence we obtain

$$\begin{aligned}
0 \le p(x_p) - p(x_q) &\le p(x_p) - p(x_q) - q(x_p) + q(x_q) \\
&\le \left| p(x_p) - q(x_p) \right| + \left| p(x_q) - q(x_q) \right| \le \sum_x \left| p(x) - q(x) \right|. \tag{40}
\end{aligned}$$

⊣

Now we can show a general result about universal predictors induced by forward estimators of conditional probabilities.

THEOREM 3.12 (Effective induced prediction I). *For a 1-universal forward estimator R, the induced predictor f_R is 1-universal if f_R is computable.*

PROOF. Let R be 1-universal forward estimator. By the definition, for every stationary ergodic measure P and all 1-P-random points

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \sum_{x_{i+1}\in\mathbb{X}} \left| P(x_{i+1}|X_1^i) - R(x_{i+1}|X_1^i) \right| = 0. \tag{41}$$

Consequently, combining this with Proposition 3.11 (prediction inequality) yields on 1-P-random points

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[P(X_{i+1} \neq f_R(X_1^i)|X_1^i) - P(X_{i+1} \neq f_P(X_1^i)|X_1^i) \right] = 0. \tag{42}$$

Now, we notice that by (29), we have on 1-P-random points

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[\mathbf{1}\{X_{i+1} \neq f_R(X_1^i)\} - P(X_{i+1} \neq f_R(X_1^i)|X_1^i) \right] = 0, \tag{43}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[\mathbf{1}\{X_{i+1} \neq f_P(X_1^i)\} - P(X_{i+1} \neq f_P(X_1^i)|X_1^i) \right] = 0. \tag{44}$$

Combining the three above observations completes the proof. ⊣

Interestingly, it suffices for a measure to be a computable almost surely universal backward estimator to yield a 1-universal forward estimator and, consequently, a 1-universal predictor. In contrast, we can easily see that a computable almost surely universal forward estimator does not necessarily induce a 1-universal predictor.

THEOREM 3.13. *There exists a computable almost surely universal forward estimator R such that the induced predictor f_R is not 1-universal.*

PROOF. Let us take $\mathbb{X} = \{0, 1\}$ and restrict ourselves to one-sided space $\mathbb{X}^\mathbb{N}$ without loss of generality. Fix a computable almost surely universal forward estimator Q. Let P_0 be the computable measure of a Bernoulli(θ) process, i.e., $P_0(x_1^n) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$, where $\theta > 1/2$ is rational. Observe that by Proposition 2.8 there exists a point $y \in \mathbb{X}^\mathbb{N}$ which is 1-P_0-random and a computable function $g : \mathbb{X}^* \to \{0, 1\}$ such that $P_0(A) = 0$ and $A = \{y\}$ for event

$$A := (\#\{i \in \mathbb{N} : g(X_1^i) = 1\} = \infty). \tag{45}$$

In other words, there is a computable method to single out some 1-P_0-random point y out of the set of sequences $\mathbb{X}^\mathbb{N}$. In particular, we can use function g to spoil measure Q on that point y while preserving the property of an almost surely universal forward estimator. We will denote the spoilt version of measure Q by R. Conditional distributions $R(X_{m+1}|X_1^m)$ will differ from $Q(X_{m+1}|X_1^m)$ for infinitely many m on point y and for finitely many m elsewhere.

Let $K(x_1^n) := \#\left\{i \leq n : g(x_1^i) = 1\right\}$. The construction of measure R proceeds by induction on the string length together with an auxiliary counter U. We let $R(x_1) := Q(x_1)$ and $U(x_1) := 0$. Suppose that $R(x_1^n)$ and $U(x_1^n)$ are defined but $R(x_1^{n+1})$ is not. If $U(x_1^n) \geq K(x_1^n)$ then we put $R(x_{n+1}|x_1^n) := Q(x_{n+1}|x_1^n)$ and $U(x_1^{n+1}) := U(x_1^n)$. Else, if $U(x_1^n) < K(x_1^n)$ then we put $R(x_{n+1}^{n+N}|x_1^n) := \prod_{i=n+1}^{n+N} \theta^{1-x_i}(1-\theta)^{x_i}$ (reverted compared to the definition of P_0!) and $U(x_{n+N}) := K(x_1^n)$ where N is the smallest number such that

$$\frac{1}{n+N}\left[\sum_{i=0}^{n-1} P_0(X_{i+1} \neq f_R(x_1^i)|X_1^i = x_1^i) + N\theta\right] \geq \frac{1}{2}. \qquad (46)$$

Such number N exists since $P_0(X_{i+1} \neq f_R(x_1^i)|X_1^i = x_1^i) > 1 - \theta$. This completes the construction of R.

The sets of 1-P-random sequences are disjoint for distinct stationary ergodic P by Theorem 2.15 (effective Birkhoff ergodic theorem). Hence $K(X_1^n)$ is bounded P-almost surely for any stationary ergodic P. Consequently, since $U(X_1^n)$ is non-decreasing then P-almost surely there exists a random number $M < \infty$ such that for all $m > M$ we have $R(X_{m+1}|X_1^m) = Q(X_{m+1}|X_1^m)$. Hence R inherits the property of an almost surely universal forward estimator from Q.

Now let us inspect what happens on y. Since $K(X_1^n)$ is unbounded on y then by the construction of R, we obtain on y that $U(X_1^n) < K(X_1^n)$ holds infinitely often and

$$\limsup_{n \to \infty} \frac{1}{n}\sum_{i=0}^{n} P_0(X_{i+1} \neq f_R(X_1^i)|X_1^i) \geq \frac{1}{2} > u_{P_0} = 1 - \theta. \qquad (47)$$

Hence predictor f_R is not 1-universal. ⊣

3.4. Predictors induced by universal measures. Following the work of Ryabko [45] (see also [46]), we can ask a natural question whether predictors induced by some universal measures in the sense of Definition 3.3, such as the PPM measure [10, 44, 47] to be discussed in Section 3.5, are also universal. Ryabko was close to demonstrate the analogous implication in the almost sure setting but did not provide the complete proof. He has shown this proposition:

THEOREM 3.14 (Theorem 3.3 in [44]). *Let R be an almost surely universal measure and P be a stationary ergodic measure. We have that*

$$\lim_{n \to \infty} E \frac{1}{n}\sum_{i=0}^{n-1} \left|P(X_{i+1}|X_0^i) - R(X_{i+1}|X_0^i)\right| = 0. \qquad (48)$$

At the first glance, condition (48) may seem close to condition (36), i.e., the universal forward estimator, which—as we have shown in Theorem 3.12—implies universality of the induced predictor. However, this average-case result is too weak for our needs as we seek the almost-sure and

effective version thereof. If we tried to derive universality of the induced predictor directly from (48), there are two problems on the way (in the following, $Y_n \geq 0$ stands for the expression under the expectation): Firstly, $\lim_{n\to\infty} \mathbf{E}\, Y_n = 0$ does not necessarily imply $\mathbf{E}\lim_{n\to\infty} Y_n = 0$ since the limit may not exist almost surely and, secondly, if $\mathbf{E}\lim_{n\to\infty} Y_n = 0$ then $Y_n = 0$ holds almost surely but this equality may fail on some 1-random points.

In this section, we will show that each 1-universal measure, under a relatively mild condition (1), satisfied by the PPM measure, is a 1-universal forward estimator and hence, in the light of the previous section, it induces a 1-universal predictor. We do not know yet whether this condition is necessary. We will circumvent Theorem 3.14 by applying Proposition 2.16 (effective Breiman ergodic theorem) and Theorem 2.17 (effective Azuma theorem). The first stage of our preparations includes two statements which can be called the effective conditional SMB theorem and the effective conditional universality.

PROPOSITION 3.15 (Effective conditional SMB theorem). *Let the alphabet be finite and let P be a stationary ergodic probability measure. On* 1-*P*-*random points we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[-\sum_{x_{i+1}\in\mathbb{X}} P(x_{i+1}|X_1^i) \log P(x_{i+1}|X_1^i) \right] = h_P. \qquad (49)$$

PROOF. Let us write the conditional entropy

$$W_i := \left[-\sum_{x_{i+1}\in\mathbb{X}} P(x_{i+1}|X_1^i) \log P(x_{i+1}|X_1^i) \right]. \qquad (50)$$

We have $0 \leq W_i \leq \log D$ with D being the cardinality of the alphabet. Moreover by Proposition 2.14 (effective Lévy law), on 1-*P*-random points there exists limit

$$\lim_{n\to\infty} W_n \circ T^{-n-1} = \left[-\sum_{x_0\in\mathbb{X}} P(x_0|X_{-\infty}^{-1}) \log P(x_0|X_{-\infty}^{-1}) \right]. \qquad (51)$$

Hence by Proposition 2.16 (effective Breiman ergodic theorem), on 1-(s, P)-random points

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} W_i = \mathbf{E}\left[-\sum_{x_0\in\mathbb{X}} P(x_0|X_{-\infty}^{-1}) \log P(x_0|X_{-\infty}^{-1}) \right] = h_P \qquad (52)$$

since $\mathbf{E}\left[-\log P(X_0|X_{-\infty}^{-1}) \right] = \lim_{n\to\infty} \left[-\log P(X_1^n) \right]/n = h_P$. ⊣

PROPOSITION 3.16 (Effective conditional universality). *Let the alphabet be finite and let P be a stationary ergodic probability measure. If measure R is* 1-*universal and satisfies*

$$-\log R(x_{n+1}|x_1^n) \leq \varepsilon_n \sqrt{n/\ln n}, \quad \lim_{n\to\infty} \varepsilon_n = 0, \qquad (53)$$

then on 1-*P-random points we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[- \sum_{x_{i+1}\in\mathbb{X}} P(x_{i+1}|X_1^i) \log R(x_{i+1}|X_1^i) \right] = h_P. \qquad (54)$$

PROOF. Let us write the conditional pointwise entropy $Z_i := -\log R(X_{i+1}|X_1^i)$. Now suppose that measure R is 1-universal and satisfies (53). Then by Theorem 2.17 (effective Azuma theorem), on 1-*P*-random points we obtain

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{E}\left(Z_i | X_1^i\right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} Z_i = \lim_{n\to\infty} \frac{1}{n} \left[-\log R(X_1^n) \right] = h_P, \tag{55}$$

which is the claim of Proposition 3.16.   ⊣

In the second stage of our preparations, we recall the famous Pinsker inequality used by Ryabko [45] to prove Theorem 3.14.

PROPOSITION 3.17 (Pinsker inequality [11]). *Let p and q be probability distributions over a countable alphabet* $\mathbb{X}$. *We have*

$$\left[\sum_{x\in\mathbb{X}} |p(x) - q(x)| \right]^2 \leq (2\ln 2) \sum_{x\in\mathbb{X}} p(x) \log \frac{p(x)}{q(x)}. \qquad (56)$$

Now we can show the main result of this section, namely, that every universal measure which satisfies a mild condition induces a universal predictor.

THEOREM 3.18 (Effective induced prediction II). *If measure R is* 1-*universal and satisfies* (53) *then it is a* 1-*universal forward estimator.*

PROOF. Let R be a 1-universal measure, whereas P be the stationary ergodic measure. By Propositions 3.15 (effective conditional SMB theorem) and 3.16 (effective conditional universality), on 1-*P*-random points we obtain

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[\sum_{x_{i+1}} P(x_{i+1}|X_1^i) \log \frac{P(x_{i+1}|X_1^i)}{R(x_{i+1}|X_1^i)} \right] = 0. \qquad (57)$$

Hence by Proposition 3.17 (Pinsker inequality), we derive on 1-*P*-random points

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left[\sum_{x_{i+1}} |P(x_{i+1}|X_1^i) - R(x_{i+1}|X_1^i)| \right]^2 = 0. \qquad (58)$$

Subsequently, the Cauchy–Schwarz inequality $\mathbf{E}\,Y^2 \geq (\mathbf{E}\,Y)^2$ yields on 1-P-random points

$$0 \geq \lim_{n\to\infty} \left[\frac{1}{n} \sum_{i=0}^{n-1} \sum_{x_{i+1}} \left| P(x_{i+1}|X_1^i) - R(x_{i+1}|X_1^i) \right| \right]^2$$

$$= \left[\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \sum_{x_{i+1}} \left| P(x_{i+1}|X_1^i) - R(x_{i+1}|X_1^i) \right| \right]^2 \geq 0. \qquad (59)$$

Consequently, R is a 1-universal forward estimator. ⊣

Combining Theorems 3.18 and 3.12, we obtain that predictor f_R is 1-universal provided measure R is 1-universal and satisfies condition (53)— if predictor f_R is computable itself. Condition (53) does not seem to have been discussed in the literature of universal prediction.

3.5. PPM measure. In this section, we will discuss the Prediction by Partial Matching (PPM) measure. The PPM measure comes in several flavors and was discovered gradually. Cleary and Witten [10] coined the name PPM, which we prefer since it is more distinctive, and considered the adaptive Markov approximations PPM_k defined roughly in Equation (63). Later, Ryabko [44, 47] considered the infinite series PPM defined in Equation (64), called it the measure R, and proved that it is a universal measure. Precisely, Ryabko used the Krichevsky–Trofimov smoothing $(+1/2)$ rather than the Laplace smoothing $(+1)$ applied in (63). This difference does not affect universality. As we will show now, the series PPM provides an example of a 1-universal measure that satisfies condition (53) and thus yields a natural 1-universal predictor.

Upon the first reading, the definition of the PPM measure may appear cumbersome but it is roughly a Bayesian mixture of all Markov chains of all orders. Its universality can be then motivated by the fact that Markov chains with rational transition probabilities are both countable and dense in the class of stationary ergodic measures [48]. Our specific definition of the PPM measure is as follows.

DEFINITION 3.19 (PPM measure). Let the alphabet be $\mathbb{X} = \{a_1, \ldots, a_D\}$, where $D \geq 2$.

Define the frequency of a substring w_1^k in a string x_1^n as

$$N(w_1^k|x_1^n) := \sum_{i=1}^{n-k+1} \mathbf{1}\left\{ x_i^{i+k-1} = w_1^k \right\}. \qquad (60)$$

Adapting the definitions by [10, 15, 44, 47], the PPM measure of order $k \geq 0$ is defined as

$$\text{PPM}_k(x_1^{k+1}) := D^{-k-1}, \qquad (61)$$

$$\text{PPM}_k(x_{n+1}|x_1^n) := \frac{N(x_{n+1-k}^{n+1}|x_1^n) + 1}{N(x_{n+1-k}^n|x_1^{n-1}) + D}, \quad n \geq k + 1, \qquad (62)$$

$$\text{PPM}_k(x_1^n) := \text{PPM}_k(x_1^{k+1}) \prod_{i=k+1}^{n} \text{PPM}_k(x_{i+1}|x_1^i). \qquad (63)$$

Subsequently, we define the total PPM measure

$$\text{PPM}(x_1^n) := \sum_{k=0}^{\infty} \left[\frac{1}{k+1} - \frac{1}{k+2} \right] \text{PPM}_k(x_1^n). \qquad (64)$$

The infinite series (64) is computable since $\text{PPM}_k(x_1^n) = D^{-n}$ for $k \geq n - 1$. The almost sure universality of the total PPM measure follows by the Stirling approximation and the Birkhoff ergodic theorem (see [15, 44, 47]). Since the Birkhoff ergodic theorem can be effectivized for 1-random points in the form of Proposition 2.15, we obtain in turn this effectivization.

THEOREM 3.20 (Effective PPM universality; cf. [44]). *The* PPM *measure is* 1-*universal*.

PROOF. As we have mentioned, computability of the PPM measure follows since series (64) can be truncated with the constant term $\text{PPM}_k(x_1^n) = D^{-n}$ for $k \geq n - 1$ and thus values $\text{PPM}(x_1^n)$ are rational.

To show 1-universality of the PPM measure, we first observe that

$$\begin{aligned}
\text{PPM}_k(X_1^n) &= D^{-k} \prod_{w_1^k} \frac{\prod_{w_{k+1}} 1 \cdot 2 \cdot \cdots \cdot N(w_1^{k+1}|x_1^n)}{D \cdot (D+1) \cdot \cdots \cdot (N(w_1^k|x_1^{n-1}) + D - 1)} \\
&= D^{-k} \prod_{w_1^k} \frac{(D-1)! \prod_{w_{k+1}} N(w_1^{k+1}|x_1^n)!}{(N(w_1^k|x_1^{n-1}) + D - 1)!}.
\end{aligned} \qquad (65)$$

In contrast, the empirical (conditional) entropy of string x_1^n of order $k \geq 0$ is defined as

$$h_k(x_1^n) := \sum_{w_1^{k+1}} \frac{N(w_1^{k+1}|x_1^n)}{n-k} \log \frac{N(w_1^k|x_1^{n-1})}{N(w_1^{k+1}|x_1^n)}. \qquad (66)$$

Using the Stirling approximation for the factorial function, the PPM measure of order $k \geq 0$ can be related to the empirical entropy. In particular, by Theorem A4 in [15], we have the bound

$$0 \leq - \log \text{PPM}_k(x_1^n) - k \log D - (n-k)h_k(x_1^n) \leq D^{k+1} \log[e^2 n]. \qquad (67)$$

Subsequently, by Proposition 2.15 (effective Birkhoff ergodic theorem), on 1-P-random points we have

$$\lim_{n \to \infty} \frac{N(w_1^{k+1}|X_1^n)}{n-k} = P(w_1^{k+1}). \qquad (68)$$

Then by (67),

$$\lim_{n\to\infty} \frac{1}{n}\left[-\log \text{PPM}_k(X_1^n)\right] = h_{k,P} := \mathbf{E}\left[-\log P(X_{k+1}|X_1^k)\right]. \quad (69)$$

Since

$$-\log \text{PPM}(x_1^n) \le 2\log(k+2) - \log \text{PPM}_k(x_1^n), \quad (70)$$

then

$$\limsup_{n\to\infty} \frac{1}{n}\left[-\log \text{PPM}(X_1^n)\right] \le \inf_{k\ge 0} h_{k,P} = h_P \quad (71)$$

on 1-P-random points, whereas the reverse inequality for the lower limit follows by Proposition 2.10 (effective Barron lemma) and Theorem 3.1 (effective SMB theorem). ⊣

Finally, we can show that predictor f_{PPM} induced by the PPM measure is 1-universal. First, we notice explicitly these bounds:

THEOREM 3.21 (PPM bounds). *We have*

$$-\log \text{PPM}(x_1^n) \le 2\log(n+1) + n\log D, \quad (72)$$

$$-\log \text{PPM}(x_{n+1}|x_1^n) \le 3\log(n+D). \quad (73)$$

PROOF. Observe that $\text{PPM}_k(x_1^n) = D^{-n}$ for $k \ge n-1$. Hence by (70), we obtain claim (72). The derivation of claim (73) is slightly longer. First, by the definition of PPM_k, we have

$$-\log \text{PPM}_k(x_{n+1}|x_1^n) \le \log\left[N(x_{n-k+1}^n|x_1^{n-1}) + D\right] \le \log(n+D). \quad (74)$$

Now let us denote

$$G := \arg\max_{k\ge 0} \text{PPM}_k(x_1^n). \quad (75)$$

We have $G \le n-1$, since $\text{PPM}_k(x_1^n) = D^{-n}$ for $k \ge n-1$. Moreover, we have a bound reverse to (70), namely

$$-\log \text{PPM}(x_1^n) \ge -\log \text{PPM}_G(x_1^n). \quad (76)$$

Combining the above with (70) yields

$$-\log \text{PPM}(x_{n+1}|x_1^n) = -\log \text{PPM}(x_1^{n+1}) + \log \text{PPM}(x_1^n)$$
$$\le 2\log(G+2) - \log \text{PPM}_G(x_1^{n+1}) + \log \text{PPM}_G(x_1^n)$$
$$= 2\log(G+2) + \log \text{PPM}_G(x_{n+1}|x_1^n) \le 3\log(n+D). \quad (77)$$

⊣

Now comes the main theorem.

THEOREM 3.22. *The predictor f_{PPM} is 1-universal.*

PROOF. Computability of the predictor f_{PPM} follows since values $\text{PPM}(x_1^n)$ are rational so the least symbol of those having the maximal conditional probability can be computed in a finite time. Consequently, the claim follows by Theorems 3.18, 3.20, and 3.21. ⊣

We think that 1-universality of the predictor f_{PPM} is quite expected and intuitive. But as we can see, the PPM measure satisfies condition (53) with a large reserve. It is an open question whether there are 1-universal measures such that conditional probabilities $R(x_{n+1}|x_1^n)$ converge to zero much faster than for the PPM measure but they still induce 1-universal predictors. It would be interesting to find such measures. Maybe they have some other desirable properties, also from a practical point of view.

REFERENCES

[1] P. H. ALGOET, *The strong law of large numbers for sequential decisions under uncertainty*. **IEEE Transactions on Information Theory**, vol. 40 (1994), no. 3, pp. 609–633.

[2] P. H. ALGOET and T. M. COVER, *A sandwich proof of the Shannon–McMillan–Breiman theorem*. **Annals of Probability**, vol. 16 (1988), pp. 899–909.

[3] K. AZUMA, *Weighted sums of certain dependent random variables*. **Tohoku Mathematical Journal. Second Series**, vol. 19 (1967), no. 3, pp. 357–367.

[4] D. H. BAILEY, *Sequential schemes for classifying and predicting ergodic processes*, Ph.D. thesis, Stanford University, 1976.

[5] A. R. BARRON, *Logically smooth density estimation*, Ph.D. thesis, Stanford University, 1985.

[6] L. BIENVENU, A. R. DAY, M. HOYRUP, I. MEZHIROV, and A. SHEN, *A constructive version of Birkhoff's ergodic theorem for Martin-Löf random points*. **Information and Computation**, vol. 210 (2012), pp. 21–30.

[7] L. BREIMAN, *The individual ergodic theorem of information theory*. **Annals of Mathematical Statistics**, vol. 28 (1957), pp. 809–811.

[8] M. CHARIKAR, E. LEHMAN, A. LEHMAN, D. LIU, R. PANIGRAHY, M. PRABHAKARAN, A. SAHAI, and A. SHELAT, *The smallest grammar problem*. **IEEE Transactions on Information Theory**, vol. 51 (2005), pp. 2554–2576.

[9] K. L. CHUNG, *A note on the ergodic theorem of information theory*. **Annals of Mathematical Statistics**, vol. 32 (1961), pp. 612–614.

[10] J. CLEARY and I. WITTEN, *Data compression using adaptive coding and partial string matching*. **IEEE Transactions on Communications**, vol. 32 (1984), pp. 396–402.

[11] I. CSISZÁR and J. KÖRNER, **Information Theory: Coding Theorems for Discrete Memoryless Systems**, Cambridge University Press, Cambridge, 2011.

[12] A. P. DAWID and V. G. VOVK, *Prequential probability*: *Principles and properties*. **Bernoulli**, vol. 5 (1999), no. 1, pp. 125–162.

[13] A. D. DAY and J. S. MILLER, *Randomness for non-computable measures*. **Transactions of the American Mathematical Society**, vol. 365 (2013), no. 7, pp. 3575–3591.

[14] Ł. DĘBOWSKI, *On the vocabulary of grammar-based codes and the logical consistency of texts*. **IEEE Transactions on Information Theory**, vol. 57 (2011), pp. 4589–4599.

[15] ———, *Is natural language a perigraphic process? The theorem about facts and words revisited*. **Entropy**, vol. 20 (2018), no. 2, p. 85.

[16] ———, *Information Theory Meets Power Laws: Stochastic Processes and Language Models*, Wiley, New York, 2021.

[17] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[18] R. G. Downey and D. R. Hirschfeldt, *Algorithmic Randomness and Complexity*, Springer, New York, 2010.

[19] R. M. Fano, *Transmission of Information*, MIT Press, Cambridge, 1961.

[20] L. Fortnow and J. H. Lutz, *Prediction and dimension*. *Journal of Computer and System Science*, vol. 70 (2005), no. 4, pp. 570–589.

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

[22] J. N. Y. Franklin and H. Towsner, *Randomness and non-ergodic systems*. *Moscow Mathematical Journal*, vol. 14 (2014), pp. 711–744.

[23] P. Gács, *Uniform test of algorithmic randomness over a general space*. *Theoretical Computer Science*, vol. 341 (2005), pp. 91–137.

[24] L. Györfi and G. Lugosi, *Strategies for sequential prediction of stationary time series*, *Modeling Uncertainty: An Examination of Its Theory, Methods, and Applications* (M. Dror, P. L'Ecuyer, and F. Szidarovszky, editors), Kluwer Academic, Dordrecht, 2001.

[25] L. Györfi, G. Lugosi, and G. Morvai, *A simple randomized algorithm for consistent sequential prediction of ergodic time series*. *IEEE Transactions on Information Theory*, vol. 45 (1999), no. 7, pp. 2642–2650.

[26] M. Hochman, *Upcrossing inequalities for stationary sequences and applications*. *Annals of Probability*, vol. 37 (2009), no. 6, pp. 2135–2149.

[27] M. Hoyrup, *The dimension of ergodic random sequences*, *29th International Symposium on Theoretical Aspects of Computer Science*, Dagstuhl Publishing, Wadern, 2012, p. 567.

[28] M. Hoyrup and C. Rojas, *Applications of effective probability theory to Martin-Löf randomness*, *International Colloquium on Automata, Languages, and Programming*, Dagstuhl Publishing, Wadern, 2009, pp. 549–561.

[29] ———, *Computability of probability measures and Martin-Löf randomness over metric spaces*. *Information and Computation*, vol. 207 (2009), no. 7, pp. 830–847.

[30] Y. Kalnishkan, M. V. Vyugin, and V. Vovk, *Generalised entropies and asymptotic complexities of languages*. *Information and Computation*, vol. 237 (2014), pp. 101–104.

[31] J. C. Kieffer and E. Yang, *Grammar-based codes*: *A new class of universal lossless source codes*. *IEEE Transactions on Information Theory*, vol. 46 (2000), pp. 737–754.

[32] L. A. Levin, *The concept of a random sequence*. *Doklady Akademii Nauk SSSR*, vol. 21 (1973), pp. 548–550.

[33] ———, *Uniform tests for randomness*. *Doklady Akademii Nauk SSSR*, vol. 227 (1976), no. 1, pp. 33–35.

[34] ———, *Randomness conservation inequalities*: *Information and independence in mathematical theories*. *Information and Control*, vol. 61 (1984), no. 1, pp. 15–37.

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

[36] B. McMillan, *Two inequalities implied by unique decipherability*. *IRE Transactions on Information Theory*, vol. 2 (1956), pp. 115–116.

[37] G. Morvai and B. Weiss, *On universal algorithms for classifying and predicting stationary processes*. *Probability Surveys*, vol. 18 (2021), pp. 77–131.

[38] M. Nakamura, *Ergodic theorems for algorithmically random sequences*, *Proceedings of the IEEE ISOC ITW on Coding and Complexity*, Institute of Electrical and Electronics Engineers, New York, 2005, pp. 147–150.

[39] S. Nandakumar, *An effective ergodic theorem and some applications*, *Proceedings of the 40th Annual Symposium on the Theory of Computing*, Association for Computing Machinery, New York, 2008, pp. 39–44.

[40] D. S. Ornstein, *Guessing the next output of a stationary process*. **Israel Journal of Mathematics**, vol. 30 (1978), no. 3, pp. 292–296.

[41] D. Osherson and S. Weinstein, *Recognizing strong random reals*. **Review of Symbolic Logic**, vol. 1 (2008), no. 1, pp. 56–63.

[42] J. Reimann, *Effectively closed sets of measures and randomness*. **Annals of Pure and Applied Logic**, vol. 156 (2008), no. 1, pp. 170–182.

[43] J. Reimann and T. A. Slaman, *Measures and their random reals*. **Transactions of the American Mathematical Society**, vol. 367 (2015), pp. 5081–5097.

[44] B. Ryabko, *Applications of Kolmogorov complexity and universal codes to nonparametric estimation of characteristics of time series*. **Fundamenta Informaticae**, vol. 83 (2008), nos. 1–2, pp. 177–196.

[45] ———, *Compression-based methods for nonparametric prediction and estimation of some characteristics of time series*. **IEEE Transactions on Information Theory**, vol. 55 (2009), no. 9, pp. 4309–4315.

[46] B. Ryabko, J. Astola, and M. Malyutov, **Compression-Based Methods of Statistical Analysis and Prediction of Time Series**, Springer, New York, 2016.

[47] B. Y. Ryabko, *Prediction of random sequences and universal coding*. **Problems of Information Transmission**, vol. 24 (1988), no. 2, pp. 87–96.

[48] D. Ryabko, *On finding predictors for arbitrary families of processes*. **Journal of Machine Learning Research**, vol. 11 (2010), pp. 581–602.

[49] M. Smorodinsky, **Ergodic Theory, Entropy**, Lecture Notes in Mathematics, vol. 214, Springer, New York, 1971.

[50] R. J. Solomonoff, *A formal theory of inductive inference, part 1 and part 2*. **Information and Control**, vol. 7 (1964), pp. 1–22, 224–254.

[51] T. Steifer, **Computable prediction of infinite binary sequences with zero–one loss**, Ph.D. thesis, Institute of Computer Science, Polish Academy of Sciences, 2020.

[52] ———, *A note on learning-theoretic characterizations of randomness and convergence*. **Review of Symbolic Logic**, (2021), pp. 1–16 (First View).

[53] J. Suzuki, *Universal prediction and universal coding*. **Systems and Computers in Japan**, vol. 34 (2003), no. 6, pp. 1–11.

[54] H. Takahashi, *On a definition of random sequences with respect to conditional probability*. **Information and Computation**, vol. 206 (2008), no. 12, pp. 1375–1382.

[55] V. V. V'yugin, *Ergodic theorems for individual random sequences*. **Theoretical Computer Science**, vol. 207 (1998), no. 2, pp. 343–361.

[56] J. Ziv and A. Lempel, *A universal algorithm for sequential data compression*. **IEEE Transactions on Information Theory**, vol. 23 (1977), pp. 337–343.

INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
01-248 WARSZAWA, POLAND
E-mail: ldebowsk@ipipan.waw.pl

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