Available online at www.sciencedirect.com
ScienceDirect

Theoretical Computer Science 382 (2007) 247–261

Theoretical Computer Science

www.elsevier.com/locate/tcs

On semimeasures predicting Martin-Löf random sequences

Marcus Hutter[a,b,*], Andrej Muchnik[c]

[a] IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
[b] RSISE/ANU/NICTA, Canberra, ACT, 0200, Australia
[c] Institute of New Technologies, 10 Nizhnyaya Radischewskaya, Moscow 109004, Russia

Abstract

Solomonoff's central result on induction is that the prediction of a universal semimeasure M converges rapidly and with probability 1 to the true sequence generating predictor μ, if the latter is computable. Hence, M is eligible as a universal sequence predictor in the case of unknown μ. Despite some nearby results and proofs in the literature, the stronger result of convergence for all (Martin-Löf) random sequences remained open. Such a convergence result would be particularly interesting and natural, since randomness can be defined in terms of M itself. We show that there are universal semimeasures M which do not converge to μ on all μ-random sequences, i.e. we give a partial negative answer to the open problem. We also provide a positive answer for some non-universal semimeasures. We define the incomputable measure D as a mixture over all computable measures and the enumerable semimeasure W as a mixture over all enumerable nearly measures. We show that W converges to D and D to μ on all random sequences. The Hellinger distance measuring closeness of two distributions plays a central role.

Keywords: Sequence prediction; Algorithmic information theory; Universal enumerable semimeasure; Mixture distributions; Predictive convergence; Martin-Löf randomness; Supermartingales; Quasimeasures

1. Introduction

> *"All difficult conjectures should be proved by reductio ad absurdum arguments. For if the proof is long and complicated enough you are bound to make a mistake somewhere and hence a contradiction will inevitably appear, and so the truth of the original conjecture is established QED".*
>
> *— Barrow's second 'law' (2004)*

A sequence prediction task is defined as to predict the next symbol x_n from an observed sequence $x = x_1 \dots x_{n-1}$. The key concept to attack general prediction problems is Occam's razor, and to a lesser extent Epicurus's principle of multiple explanations. The former/latter may be interpreted as to keep the simplest/all theories consistent with the observations $x_1 \dots x_{n-1}$ and to use these theories to predict x_n. Solomonoff [13,14] formalized and combined

[*] Corresponding author at: RSISE/ANU/NICTA, Canberra, ACT, 0200, Australia.
E-mail addresses: marcus@hutter1.net (M. Hutter), muchnik@lpcs.math.msu.su (A. Muchnik).
URL: http://www.hutter1.net (M. Hutter).

0304-3975/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2007.03.040

both principles in his universal a priori semimeasure M which assigns high/low probability to simple/complex environments x, hence implementing Occam and Epicurus. Formally it can be represented as a mixture of all enumerable semimeasures. An abstract characterization of M by Levin [17] is that M is a universal enumerable semimeasure in the sense that it multiplicatively dominates all enumerable semimeasures.

Solomonoff's [14] central result is that if the probability $\mu(x_n|x_1 \ldots x_{n-1})$ of observing x_n at time n, given past observations $x_1 \ldots x_{n-1}$ is a computable function, then the universal predictor $M_n := M(x_n|x_1 \ldots x_{n-1})$ converges (rapidly!) *with μ-probability* 1 (w.p.1) for $n \to \infty$ to the optimal/true/informed predictor $\mu_n := \mu(x_n|x_1 \ldots x_{n-1})$, hence M represents a universal predictor in the case of unknown "true" distribution μ. Convergence of M_n to μ_n w.p.1 tells us that M_n is close to μ_n for sufficiently large n for almost all sequences $x_1x_2 \ldots$. It says nothing about whether convergence is true for any *particular* sequence (of measure 0).

Martin-Löf (M.L.) randomness is the standard notion for randomness of individual sequences [9,8]. A M.L.-random sequence passes *all* thinkable effective randomness tests, e.g. the law of large numbers, the law of the iterated logarithm, etc. In particular, the set of all μ-random sequences has μ-measure 1. It is natural to ask whether M_n converges to μ_n (in difference or ratio) individually for all M.L.-random sequences. Clearly, Solomonoff's result shows that convergence may at most fail for a set of sequences with μ-measure zero. A convergence result for M.L.-random sequences would be particularly interesting and natural in this context, since M.L.-randomness can be defined in terms of M itself [7]. Despite several attempts to solve this problem [16,15,3], it remained open [4].

In this paper we construct an M.L.-random sequence and show the existence of a universal semimeasure which does not converge on this sequence, hence answering the open question negatively for some M. It remains open whether there exist (other) universal semimeasures, probably with particularly interesting additional structure and properties, for which M.L.-convergence holds. The main positive contribution of this work is the construction of a non-universal enumerable semimeasure W which M.L.-converges to μ as desired. As an intermediate step we consider the incomputable measure $\hat{D}$, defined as a mixture over all computable measures. We show M.L.-convergence of predictor W to $\hat{D}$ and of $\hat{D}$ to μ. The Hellinger distance measuring closeness of two predictive distributions plays a central role in this work.

The paper is organized as follows: In Section 2 we give basic notation and results (for strings, numbers, sets, functions, asymptotics, computability concepts, prefix Kolmogorov complexity), and define and discuss the concepts of (universal) (enumerable) (semi)measures. Section 3 summarizes Solomonoff's and Gács' results on predictive convergence of M to μ with probability 1. Both results can be derived from a bound on the expected Hellinger sum. We present an improved bound on the expected exponentiated Hellinger sum, which implies very strong assertions on the convergence rate. In Section 4 we investigate whether convergence for all Martin-Löf random sequences hold. We construct a μ-M.L.-random sequence on which some universal semimeasures M do not converge to μ. We give a non-constructive and a constructive proof of different virtue. In Section 5 we present our main positive result. We derive a finite bound on the Hellinger sum between μ and $\hat{D}$, which is exponential in the randomness deficiency of the sequence and double exponential in the complexity of μ. This implies that the predictor $\hat{D}$ M.L.-converges to μ. Finally, in Section 6 we show that W is non-universal and asymptotically M.L.-converges to $\hat{D}$, and summarize the computability, measure, and dominance properties of M, D, $\hat{D}$, and W. Section 7 contains discussion and outlook.

2. Notation & universal semimeasures *M*

Strings. Let $i, k, n, t \in \mathbb{N} = \{1, 2, 3, \ldots\}$ be natural numbers, $x, y, z \in \mathcal{X}^* = \bigcup_{n=0}^{\infty} \mathcal{X}^n$ be finite strings of symbols over finite alphabet $\mathcal{X} \ni a, b$. We write xy for the concatenation of string x with y. We denote strings x of length $\ell(x) = n$ by $x = x_1x_2 \ldots x_n \in \mathcal{X}^n$ with $x_t \in \mathcal{X}$ and further abbreviate $x_{k:n} := x_kx_{k+1} \ldots x_{n-1}x_n$ for $k \leq n$, and $x_{<n} := x_1 \ldots x_{n-1}$, and $\epsilon = x_{<1} = x_{n+1:n} \in \mathcal{X}^0 = \{\epsilon\}$ for the empty string. Let $\omega = x_{1:\infty} \in \mathcal{X}^\infty$ be a generic and $\alpha \in \mathcal{X}^\infty$ a specific infinite sequence. For a given sequence $x_{1:\infty}$ we say that x_t is on-sequence and $\bar{x}_t \neq x_t$ is off-sequence. x'_t may be on- or off-sequence. We identify strings with natural numbers (including zero, $\mathcal{X}^* \cong \mathbb{N} \cup \{0\}$).
Sets and functions. $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{R}_+ := [0, \infty)$ are the sets of fractional, real, and non-negative real numbers, respectively. $\#\mathcal{S}$ denotes the number of elements in set $\mathcal{S}$, $\ln()$ the natural and $\log()$ the binary logarithm.

Asymptotics. We abbreviate $\lim_{n\to\infty}[f(n) - g(n)] = 0$ by $f(n) \overset{n\to\infty}{\longrightarrow} g(n)$ and say f converges to g, without implying that $\lim_{n\to\infty} g(n)$ itself exists. We write $f(x) \overset{\times}{\leq} g(x)$ for $f(x) = O(g(x))$ and $f(x) \overset{+}{\leq} g(x)$ for $f(x) \leq g(x) + O(1)$.

Computability. A function $f : \mathcal{S} \to \mathbb{R} \cup \{\infty\}$ is said to be enumerable (or lower semicomputable) if the set $\{(x, y) : y < f(x), x \in \mathcal{S}, y \in \mathbb{Q}\}$ is recursively enumerable. f is co-enumerable (or upper semicomputable) if $[-f]$ is enumerable. f is computable (or estimable or recursive) if f and $[-f]$ are enumerable. f is approximable (or limit-computable) if there is a computable function $g : \S \times \mathbb{N} \to \mathbb{R}$ with $\lim_{n\to\infty} g(x, n) = f(x)$.
Complexity. The conditional prefix (Kolmogorov) complexity $K(x|y) := \min\{\ell(p) : U(y, p) = x \text{ halts}\}$ is the length of the shortest binary program $p \in \{0, 1\}^*$ on a universal prefix Turing machine U with output $x \in \mathcal{X}^*$ and input $y \in \mathcal{X}^*$ [8]. $K(x) := K(x|\epsilon)$. For non-string objects o we define $K(o) := K(\langle o\rangle)$, where $\langle o\rangle \in \mathcal{X}^*$ is some standard code for o. In particular, if $(f_i)_{i=1}^{\infty}$ is an enumeration of all enumerable functions, we define $K(f_i) \overset{+}{=} K(i)$. We only need the following elementary properties: The co-enumerability of K, the upper bounds $K(x|\ell(x)) \overset{+}{\leq} \ell(x) \log|\mathcal{X}|$ and $K(n) \overset{+}{\leq} 2\log n$, and $K(x|y) \overset{+}{\leq} K(x)$, subadditivity $K(x) \overset{+}{\leq} K(x, y) \overset{+}{\leq} K(y) + K(x|y)$, and information non-increase $K(f(x)) \overset{+}{\leq} K(x) + K(f)$ for recursive $f : \mathcal{X}^* \to \mathcal{X}^*$.

We need the concepts of (universal) (semi)measures for strings [17].

Definition 1. (*(Semi)measures*). We call $\nu : \mathcal{X}^* \to [0, 1]$ a semimeasure if $\nu(x) \geq \sum_{a\in\mathcal{X}} \nu(xa)\,\forall x \in \mathcal{X}^*$, and a (probability) measure if equality holds and $\nu(\epsilon) = 1$. $\nu(x)$ denotes the ν-probability that a sequence starts with string x. Further, $\nu(a|x) := \frac{\nu(xa)}{\nu(x)}$ is the predictive ν-probability that the next symbol is $a \in \mathcal{X}$, given sequence $x \in \mathcal{X}^*$.

Definition 2 (*Universal Semimeasures M*). A semimeasure M is called a universal element of a class of semimeasures $\mathcal{M}$, if it multiplicatively dominates all members in the sense that

$$M \in \mathcal{M} \text{ and } \forall\nu \in \mathcal{M}\ \exists w_\nu > 0 : M(x) \geq w_\nu \cdot \nu(x)\ \forall x \in \mathcal{X}^*.$$

From now on we consider the (in a sense) largest class $\mathcal{M}$ which is relevant from a constructive point of view (but see [11,12,3] for even larger constructive classes), namely the class of *all* semimeasures, which can be enumerated (=effectively be approximated) from below:

$$\mathcal{M} := \text{ class of all enumerable semimeasures.} \tag{1}$$

Solomonoff [13, Eq. (7)] defined the universal predictor $M(y|x) = M(xy)/M(x)$ with $M(x)$ defined as the probability that the output of a universal monotone Turing machine starts with x when provided with fair coin flips on the input tape. Levin [17] has shown that this M is a universal enumerable semimeasure. Another possible definition of M is as a (Bayes) mixture [13,17,14,8,3,6]: $\tilde{M}(x) = \sum_{\nu\in\mathcal{M}} 2^{-K(\nu)}\nu(x)$, where $K(\nu)$ is the length of the shortest program computing function ν. Levin [17] has shown that the class of *all* enumerable semimeasures is enumerable (with repetitions), hence $\tilde{M}$ is enumerable, since K is co-enumerable. Hence $\tilde{M} \in \mathcal{M}$, which implies

$$M(x) \geq w_{\tilde{M}}\tilde{M}(x) \geq w_{\tilde{M}}2^{-K(\nu)}\nu(x) = w'_\nu\nu(x), \quad \text{where} \quad w'_\nu \overset{\times}{=} 2^{-K(\nu)}. \tag{2}$$

Up to a multiplicative constant, M assigns higher probability to all x than any other enumerable semimeasure. All M have the same very slowly decreasing (in ν) domination constants w'_ν, essentially because $M \in \mathcal{M}$. We drop the prime from w'_ν in the following. The mixture definition $\tilde{M}$ immediately generalizes to arbitrary weighted sums of (semi)measures over countable classes other than $\mathcal{M}$, but the class may not contain the mixture, and the domination constants may be rapidly decreasing. We will exploit this for the construction of the non-universal semimeasure W in Sections 5 and 6.

3. Predictive convergence with probability 1

The following convergence results for M are well known [14,8,2,6].

Theorem 3 (*Convergence of M to μ w.p.1*). *For any universal semimeasure M and any computable measure μ it holds:*

$$M(x'_n|x_{<n}) \to \mu(x'_n|x_{<n}) \textit{ for any } x'_n \textit{ and } \frac{M(x_n|x_{<n})}{\mu(x_n|x_{<n})} \to 1, \textit{ both w.p.1 for } n \to \infty.$$

The first convergence in difference is Solomonoff's [14] celebrated convergence result. The second convergence in ratio has first been derived by Gács [8]. Note the subtle difference between the two convergence results. For *any* sequence $x'_{1:\infty}$ (possibly constant and not necessarily random), $M(x'_n|x_{<n}) - \mu(x'_n|x_{<n})$ converges to zero w.p.1 (referring to $x_{1:\infty}$), but no statement is possible for $M(x'_n|x_{<n})/\mu(x'_n|x_{<n})$, since $\liminf \mu(x'_n|x_{<n})$ could be zero. On

the other hand, if we stay *on*-sequence ($x'_{1:\infty} = x_{1:\infty}$), we have $M(x_n|x_{<n})/\mu(x_n|x_{<n}) \to 1$ (whether $\inf \mu(x_n|x_{<n})$ tends to zero or not does not matter). Indeed, it is easy to give an example where $M(x'_n|x_{<n})/\mu(x'_n|x_{<n})$ diverges. For $\mu(1|x_{<n}) = 1 - \mu(0|x_{<n}) = \frac{1}{2}n^{-3}$ we get $\mu(0_{1:n}) = \prod_{t=1}^{n}(1 - \frac{1}{2}t^{-3}) \overset{n\to\infty}{\longrightarrow} c = 0.450\ldots > 0$, i.e. $0_{1:\infty}$ is μ-random. On the other hand, one can show that $M(0_{<n}) = O(1)$ and $M(0_{<n}1) \overset{\times}{=} 2^{-K(n)}$, which implies $\frac{M(1|0_{<n})}{\mu(1|0_{<n})} \overset{\times}{=} n^3 \cdot 2^{-K(n)} \overset{\times}{\geq} n \to \infty$ for $n \to \infty$ ($K(n) \overset{+}{\leq} 2\log n$).

Theorem 3 follows from (the discussion after) Lemma 4 due to $M(x) \geq w_\mu \mu(x)$. Actually the Lemma strengthens and generalizes Theorem 3. In the following we denote expectations w.r.t. measure ρ by $\mathbf{E}_\rho$, i.e. for a function $f : \mathcal{X}^n \to \mathbb{R}$, $\mathbf{E}_\rho[f] = \sum'_{x_{1:n}} \rho(x_{1:n}) f(x_{1:n})$, where $\sum'$ sums over all $x_{1:n}$ for which $\rho(x_{1:n}) \neq 0$. Using $\sum'$ instead $\sum$ is (only) important for partial functions f undefined on a set of ρ-measure zero. Similarly $\mathbf{P}_\rho$ denotes the ρ-probability.

Lemma 4 (*Expected Bounds on Hellinger Sum*). *Let μ be a measure and ν be a semimeasure with $\nu(x) \geq w \cdot \mu(x)$ $\forall x$. Then the following bounds on the Hellinger distance $h_t(\nu, \mu|\omega_{<t}) := \sum_{a\in\mathcal{X}}(\sqrt{\nu(a|\omega_{<t})} - \sqrt{\mu(a|\omega_{<t})})^2$ hold:*

$$\sum_{t=1}^{\infty} \mathbf{E}\left[\left(\sqrt{\tfrac{\nu(\omega_t|\omega_{<t})}{\mu(\omega_t|\omega_{<t})}} - 1\right)^2\right] \overset{\text{(i)}}{\leq} \sum_{t=1}^{\infty} \mathbf{E}[h_t] \overset{\text{(ii)}}{\leq} 2\ln\left\{\mathbf{E}\left[\exp\left(\frac{1}{2}\sum_{t=1}^{\infty} h_t\right)\right]\right\} \overset{\text{(iii)}}{\leq} \ln w^{-1}$$

where $\mathbf{E}$ *here and later means expectation w.r.t.* μ.

The $\ln w^{-1}$-bounds on the first and second expression have first been derived in [2], the second being a variation of Solomonoff's bound $\sum_n \mathbf{E}[(\nu(0|x_{<n}) - \mu(0|x_{<n}))^2] \leq \frac{1}{2}\ln w^{-1}$. If sequence $x_1 x_2 \ldots$ is sampled from the probability measure μ, these bounds imply

$$\nu(x'_n|x_{<n}) \to \mu(x'_n|x_{<n}) \text{ for any } x'_n \text{ and } \tfrac{\nu(x_n|x_{<n})}{\mu(x_n|x_{<n})} \to 1, \text{ both w.p.1 for } n \to \infty,$$

where w.p.1 stands here and in the following for 'with μ-probability 1'.

Convergence is "fast" in the following sense: The second bound ($\sum_t \mathbf{E}[h_t] \leq \ln w^{-1}$) implies that the expected number of times t in which $h_t \geq \varepsilon$ is finite and bounded by $\frac{1}{\varepsilon}\ln w^{-1}$. The new third bound represents a significant improvement. It implies by means of a Markov inequality that the probability of even only marginally exceeding this number is extremely small, and that $\sum_t h_t$ is very unlikely to exceed $\ln w^{-1}$ by much. More precisely:

$$\mathbf{P}\left[\#\{t : h_t \geq \varepsilon\} \geq \tfrac{1}{\varepsilon}(\ln w^{-1} + c)\right] \leq \mathbf{P}\left[\sum_t h_t \geq \ln w^{-1} + c\right]$$
$$= \mathbf{P}\left[\exp\left(\tfrac{1}{2}\sum_t h_t\right) \geq \mathrm{e}^{c/2} w^{-1/2}\right] \leq \sqrt{w}\,\mathbf{E}\left[\exp\left(\tfrac{1}{2}\sum_t h_t\right)\right]\mathrm{e}^{-c/2} \leq \mathrm{e}^{-c/2}.$$

Proof. We use the abbreviations $\rho_t = \rho(x_t|x_{<t})$ and $\rho_{1:n} = \rho_1 \cdots \rho_n = \rho(x_{1:n})$ for $\rho \in \{\mu, \nu, R, N, \ldots\}$ and $h_t = \sum_{x_t}(\sqrt{\nu_t} - \sqrt{\mu_t})^2$.

(i) follows from

$$\mathbf{E}[(\sqrt{\tfrac{\nu_t}{\mu_t}} - 1)^2|x_{<t}] \equiv \sum_{x_t:\mu_t\neq 0} \mu_t(\sqrt{\tfrac{\nu_t}{\mu_t}} - 1)^2 = \sum_{x_t:\mu_t\neq 0} (\sqrt{\nu_t} - \sqrt{\mu_t})^2 \leq h_t$$

by taking the expectation $\mathbf{E}[]$ and sum $\sum_{t=1}^{\infty}$.

(ii) follows from Jensen's inequality $\exp(\mathbf{E}[f]) \leq \mathbf{E}[\exp(f)]$ for $f = \frac{1}{2}\sum_t h_t$.

(iii) We exploit a construction used in [16, Thm.1]. For discrete (semi)measures p and q with $\sum_i p_i = 1$ and $\sum_i q_i \leq 1$ it holds:

$$\sum_i \sqrt{p_i q_i} \leq 1 - \frac{1}{2}\sum_i(\sqrt{p_i} - \sqrt{q_i})^2 \leq \exp\left[-\frac{1}{2}\sum_i(\sqrt{p_i} - \sqrt{q_i})^2\right]. \tag{3}$$

The first inequality is obvious after multiplying out the second expression. The second inequality follows from $1-x \leq e^{-x}$. Vovk [16] defined a measure $R_t := \sqrt{\mu_t \nu_t}/N_t$ with normalization $N_t := \sum_{x_t} \sqrt{\mu_t \nu_t}$. Applying (3) for measure μ and semimeasure ν we get $N_t \leq \exp(-\frac{1}{2}h_t)$. Together with $\nu(x) \geq w \cdot \mu(x)\ \forall x$ this implies

$$\prod_{t=1}^{n} R_t = \prod_{t=1}^{n} \frac{\sqrt{\mu_t \nu_t}}{N_t} = \frac{\sqrt{\mu_{1:n}\nu_{1:n}}}{N_{1:n}} = \mu_{1:n}\sqrt{\frac{\nu_{1:n}}{\mu_{1:n}}} N_{1:n}^{-1} \geq \mu_{1:n}\sqrt{w}\exp\left(\frac{1}{2}\sum_{t=1}^{n} h_t\right).$$

Summing over $x_{1:n}$ and exploiting $\sum_{x_t} R_t = 1$ we get $1 \geq \sqrt{w}\mathbf{E}[\exp(\frac{1}{2}\sum_t h_t)]$, which proves (iii).

The bound and proof may be generalized to $1 \geq w^\kappa \mathbf{E}[\exp(\frac{1}{2}\sum_t \sum_{x_t}(\nu_t^\kappa - \mu_t^\kappa)^{1/\kappa})]$ with $0 \leq \kappa \leq \frac{1}{2}$ by defining $R_t = \mu_t^{1-\kappa}\nu_t^\kappa/N_t$ with $N_t = \sum_{x_t}\mu_t^{1-\kappa}\nu_t^\kappa$ and exploiting $\sum_i p_i^{1-\kappa}q_i^\kappa \leq \exp(-\frac{1}{2}\sum_i (p_i^\kappa - q_i^\kappa)^{1/\kappa})$. □

One can show that the constant $\frac{1}{2}$ in Lemma 4 can essentially not be improved. Increasing it to a constant $\alpha > 1$ makes the expression infinite for some (Bernoulli) distribution μ (however we choose ν). For $\nu = M$ the expression can become already infinite for $\alpha > \frac{1}{2}$ and some computable measure μ.

4. Non-convergence in Martin-Löf sense

Convergence of $M(x_n|x_{<n})$ to $\mu(x_n|x_{<n})$ with μ-probability 1 tells us that $M(x_n|x_{<n})$ is close to $\mu(x_n|x_{<n})$ for sufficiently large n on 'most' sequences $x_{1:\infty}$. It says nothing whether convergence is true for any *particular* sequence (of measure 0). Martin-Löf randomness can be used to capture convergence properties for individual sequences. Martin-Löf randomness is a very important and default concept of randomness of individual sequences, which is closely related to Kolmogorov complexity and Solomonoff's universal semimeasure M. Levin gave a characterization equivalent to Martin-Löf's original definition [7]:

Definition 5 (*Martin-Löf Random Sequences*). A sequence $\omega = \omega_{1:\infty}$ is μ-Martin-Löf random (μ.M.L.) iff there is a constant $c < \infty$ such that $M(\omega_{1:n}) \leq c \cdot \mu(\omega_{1:n})$ for all n. Moreover, $d_\mu(\omega) := \sup_n\{\log\frac{M(\omega_{1:n})}{\mu(\omega_{1:n})}\} \leq \log c$ is called the randomness deficiency of ω.

One can show that an M.L.-random sequence $x_{1:\infty}$ passes *all* thinkable effective randomness tests, e.g. the law of large numbers, the law of the iterated logarithm, etc. In particular, the set of all μ.M.L.-random sequences has μ-measure 1.

The open question we study in this section is whether M converges to μ (in difference or ratio) individually for all Martin-Löf random sequences. Clearly, Theorem 3 implies that convergence μ.M.L. may at most fail for a set of sequences with μ-measure zero. A convergence M.L. result would be particularly interesting and natural for M, since M.L.-randomness can be defined in terms of M itself (Definition 5).

The state of the art regarding this problem may be summarized as follows: [16] contains a (non-improvable?) result which is slightly too weak to imply M.L.-convergence, [8, Thm. 5.2.2] and [15, Thm. 10] contain an erroneous proof for M.L.-convergence, and [3] proves a theorem indicating that the answer may be hard and subtle (see [3] for details).

The main contribution of this section is a partial answer to this question. We show that M.L.-convergence fails at least for some universal semimeasures:

Theorem 6 (*Universal Semimeasure Non-Convergence*). *There exists a universal semimeasure M and a computable measure μ and a μ.M.L.-random sequence α, such that*

$$M(\alpha_n|\alpha_{<n}) \not\to \mu(\alpha_n|\alpha_{<n}) \quad \textit{for} \quad n \to \infty.$$

This implies that also M_n/μ_n does not converge (since $\mu_n \leq 1$ is bounded). We do not know whether Theorem 6 holds for *all* universal semimeasures. For the proof we need the concept of supermartingales. We only define it for binary alphabet and uniform measure $\mu(x) = \lambda(x) := 2^{-\ell(x)}$ for which we need it.

Definition 7 (*Supermartingale*).

$$m : \{0,1\}^* \to \mathbb{R} \text{ is a supermartingale } :\Leftrightarrow m(x) \geq \tfrac{1}{2}[m(x0)+m(x1)] \text{ for all } x \in \{0,1\}^*.$$

If ν is a (enumerable) semimeasure, then $m := \nu/\lambda$ is a (enumerable) supermartingale. We prove the following theorem, which will imply Theorem 6.

Lemma 8 (*Supermartingale Non-Convergence*). *For the M.L.-random sequence α defined in (4) and the enumerable supermartingale r defined in Lemma 9 and for any $\eta, \eta' \in \mathbb{R}$ and any on α bounded supermartingale R, i.e. $0 < \varepsilon < R(\alpha_{1:n}) < c < \infty \, \forall n$, it holds that*

$$\left| \frac{R(\alpha_{1:n})}{R(\alpha_{<n})} - \eta \right| > \delta \quad \textit{or} \quad \left| \frac{R'(\alpha_{1:n})}{R'(\alpha_{<n})} - \eta' \right| > \delta$$

(or both) for a non-vanishing fraction of n, where supermartingale $R' := \frac{1}{2}(R + r)$ and some $\delta > 0$.

Proof. We define a sequence α, which, in a sense, is the lexicographically first (or equivalently leftmost in the tree of sequences) λ.M.L.-random sequence. Formally we define α, inductively in $n = 1, 2, 3, \ldots$ by

$$\alpha_n = 0 \text{ if } M(\alpha_{<n}0) \leq 2^{-n}, \text{ and } \alpha_n = 1 \text{ else.} \tag{4}$$

We know that $M(\epsilon) \leq 1$ and $M(\alpha_{<n}0) \leq 2^{-n}$ if $\alpha_n = 0$. Inductively, assuming $M(\alpha_{<n}) \leq 2^{-n+1}$ for $\alpha_n = 1$ we have $2^{-n+1} \geq M(\alpha_{<n}) \geq M(\alpha_{<n}0) + M(\alpha_{<n}1) \geq 2^{-n} + M(\alpha_{<n}1)$ since M is a semimeasure, hence $M(\alpha_{<n}1) \leq 2^{-n}$. Hence[1]

$$M(\alpha_{1:n}) \leq 2^{-n} \equiv \lambda(\alpha_{1:n}) \, \forall n, \text{ i.e. } \alpha \text{ is } \lambda\text{.M.L.-random.} \tag{5}$$

With R and r, also $R' := \frac{1}{2}(R + r) > 0$ is a supermartingale. We prove that the theorem holds for infinitely many n. It is easy to refine the proof to a non-vanishing fraction of n's. Assume that $\frac{R(\alpha_{1:n})}{R(\alpha_{<n})} \to \eta$ for $n \to \infty$ (otherwise we are done). $\eta > 1$ implies $R \to \infty$, $\eta < 1$ implies $R \to 0$. Since R is bounded, η must be 1, hence for sufficiently large n_0 we have $|R(\alpha_{1:n}) - R(\alpha_{<n})| < \varepsilon$ for all $n \geq n_0$.

Assume $r \in \{0, \frac{1}{2}, 1\}$ and $r(\alpha_{1:n}) = \frac{1}{2}$ for infinitely many n and $r(\alpha_{1:n}) = 1$ for infinitely many n (e.g. take r as defined in Lemma 9). Since R stabilizes and r oscillates, R' cannot converge. Formally, for (the infinitely many) $n \geq n_0$ for which $r(\alpha_{<n}) = \frac{1}{2}$ and $r(\alpha_{1:n}) = 1$ we have

$$\frac{R'(\alpha_{1:n})}{R'(\alpha_{<n})} - 1 \equiv \frac{R(\alpha_{1:n}) - R(\alpha_{<n}) + r(\alpha_{1:n}) - r(\alpha_{<n})}{R(\alpha_{<n}) + r(\alpha_{<n})} \geq \frac{-\varepsilon + \frac{1}{2}}{c + \frac{1}{2}} \geq \delta \; > \; 0$$

for sufficiently small ε and δ. Similarly for (the infinitely many) $n \geq n_0$ for which $r(\alpha_{<n}) = 1$ and $r(\alpha_{1:n}) = \frac{1}{2}$ we have

$$1 - \frac{R'(\alpha_{1:n})}{R'(\alpha_{<n})} \equiv \frac{R(\alpha_{<n}) - R(\alpha_{1:n}) + r(\alpha_{<n}) - r(\alpha_{1:n})}{R(\alpha_{<n}) + r(\alpha_{<n})} \geq \frac{-\varepsilon + \frac{1}{2}}{c + 1} \geq \delta \; > \; 0.$$

This shows that Lemma 8 holds for infinitely many n. If we define r zero off-sequence, i.e. $r(x) = 0$ for $x \neq \alpha_{1:\ell(x)}$, then r is a supermartingale, but a non-enumerable one, since α is not computable. In the next lemma we define an enumerable supermartingale r, which completes the proof of Lemma 8. Finally note that we could have defined $R' = \frac{R + \gamma r}{1 + \gamma}$ with arbitrarily small $\gamma > 0$, showing that already a small contamination can destroy convergence. This is no longer true for the constructive proof below. □

Lemma 9 (*Enumerable Supermartingale*). *Let M^t with $t = 1, 2, 3, \ldots$ be computable approximations of M, which enumerate M, i.e. $M^t(x) \nearrow M(x)$ for $t \to \infty$. For each t define recursively a sequence α^t similarly to (4) as $\alpha^t_n = 0$ if $M^t(\alpha^t_{<n}0) \leq 2^{-n}$ and $\alpha^t_n = 1$ else. For even $\ell(x)$ we define $r(x) = 1$ if $\exists t, n : x = \alpha^t_{<n}$ and $r(x) = 0$ else. For odd $\ell(x)$ we define $r(x) = \frac{1}{2}[r(x0) + r(x1)]$. r is an enumerable supermartingale with $r(\alpha_{1:n})$ being 1 and $\frac{1}{2}$ for a non-vanishing fraction of n's, where $\alpha = \lim_{t \to \infty} \alpha^t$ ($\alpha^t \nearrow \alpha$ lexicographically increasing).*

The idea behind the definition of r is to define $r(\alpha_{<n}) = 1$ for odd n and if possible $\frac{1}{2}$ for even n. The following possibilities exist for the local part of the sequence tree:

$$\begin{array}{c} r(x) \\ \wedge \\ r(x0) \quad r(x1) \end{array} = \begin{array}{c} 0 \\ \wedge \\ 0 \; 0 \end{array}, \ell(x) \text{ odd } \begin{array}{c} 1/2 \\ \wedge \\ 1 \; 0 \end{array} \text{ or } \begin{array}{c} 1/2 \\ \wedge \\ 0 \; 1 \end{array} \text{ or } \begin{array}{c} 1 \\ \wedge \\ 1 \; 1 \end{array}, \text{ and } \ell(x) \text{ even } \begin{array}{c} 1 \\ \wedge \\ 1/2 \; 0 \end{array} \text{ or } \begin{array}{c} 1 \\ \wedge \\ 0 \; 1/2 \end{array} \text{ or } \begin{array}{c} 1 \\ \wedge \\ 1/2 \; 1/2 \end{array},$$

all respecting the supermartingale property. The formal proof goes as follows:

[1] Alternatively we may define $\alpha_n = 0$ if $M(0|\alpha_{<t}) \leq \frac{1}{2}$ and $\alpha_n = 1$ else.

Proof. r is enumerable, since $\alpha^t_{<n}$ is computable. Further, $0 \leq r(x) \leq 1\,\forall x$. For odd $\ell(x)$ the supermartingale property $r(x) \geq \frac{1}{2}[r(x0) + r(x1)]$ is obviously satisfied. For even $\ell(x)$ and $x = \alpha^t_{<n}$ for some t we have $r(x) = 1 = \frac{1}{2}[1 + 1] \geq \frac{1}{2}[r(x0) + r(x1)]$. Even $\ell(x)$ and $x \neq \alpha^t_{<n}\,\forall t$ implies $xy \neq \alpha^t_{1:\ell(xy)}\,\forall t, y$, hence $r(x) = 0 = \frac{1}{2}[0 + 0] = \frac{1}{2}[r(x0) + r(x1)]$. This shows that r is a supermartingale.

Since M^t is monotone increasing, α^t is also monotone increasing w.r.t. to lexicographical ordering on $\{0, 1\}^\infty$. Hence $\alpha^t_{1:n}$ converges to $\alpha_{1:n}$ for $t \to \infty$, and even $\alpha^t_{1:n} = \alpha_{1:n}\,\forall t \geq t_n$ and sufficiently large (n-dependent) t_n. This implies $r(\alpha_{<n}) = r(\alpha^{t_n}_{<n}) = 1$ for odd n. We know that $\alpha_n = 0$ for a non-vanishing fraction of (even) n, since α is random. For such n, $\alpha^t_n = 0\,\forall t$, hence $r(\alpha_{<n}) = r(\alpha^{t_n}_{<n}) = \frac{1}{2}[r(\alpha^{t_n}_{<n}0) + r(\alpha^{t_n}_{<n}1)] = \frac{1}{2}[1 + 0] = \frac{1}{2}$. This shows that $r(\alpha_{<n}) = 1$ $(\frac{1}{2})$ for a non-vanishing fraction of n, namely the odd ones (the even ones with $\alpha_n = 0$). □

Non-constructive Proof of Theorem 6. Use Lemma 8 with $R := M/\lambda$, $R' := M'/\lambda$, $r =: q/\lambda$, hence q is an enumerable semimeasure, hence with M, also $M' = \frac{1}{2}(M + q)$ is a universal semimeasure. $R(\alpha_{1:n}) \leq 1$ from (5) and $R(x) \geq c > 0$ from universality of M and computability of λ show that the conditions of Lemma 8 are satisfied. Hence $R^{(\prime)}(\alpha_{1:n})/R^{(\prime)}(\alpha_{<n}) \equiv M^{(\prime)}(\alpha_n|\alpha_{<n})/\lambda(\alpha_n|\alpha_{<n}) \not\to 1$. Multiplying this by $\lambda_n = \mu_n = \frac{1}{2}$ completes the proof. □

The proof of Theorem 6 is non-constructive. Either M or M' (or both) do not converge, but we do not know which one. Below we give an alternative proof which is constructive. The idea is to construct an enumerable (semi)measure ν such that ν dominates M on α, but $\nu(\alpha_n|\alpha_{<n}) \not\to \frac{1}{2}$. Then we mix M to ν to make ν universal, but with larger contribution from ν, in order to preserve non-convergence.

Constructive Proof of Theorem 6. We define an enumerable semimeasure ν as follows:

$$\nu^t(x) := \begin{cases} 2^{-t} & \text{if} \quad \ell(x) = t \quad \text{and} \quad x < \alpha^t_{1:t} \\ 0 & \text{if} \quad \ell(x) = t \quad \text{and} \quad x \geq \alpha^t_{1:t} \\ 0 & \text{if} \quad \ell(x) > t \\ \nu^t(x0) + \nu^t(x1) & \text{if} \quad \ell(x) < t \end{cases} \tag{6}$$

where $<$ is the lexicographical ordering on sequences, and α^t has been defined in Lemma 9. ν^t is a semimeasure, and with α^t also ν^t is computable and monotone increasing in t, hence $\nu := \lim_{t\to\infty} \nu^t$ is an enumerable semimeasure (indeed, $\frac{\nu(x)}{\nu(\epsilon)}$ is a measure). We could have defined a ν_{tn} by replacing $\alpha^t_{1:t}$ with $\alpha^n_{1:t}$ in (6). Since ν_{tn} is monotone increasing in t and n, any order of $t, n \to \infty$ leads to ν, so we have chosen arbitrarily $t = n$. By induction (starting from $\ell(x) = t$) it follows that

$$\nu^t(x) = 2^{-\ell(x)} \quad \text{if} \quad x < \alpha^t_{1:\ell(x)} \quad \text{and} \quad \ell(x) \leq t, \qquad \nu^t(x) = 0 \quad \text{if} \quad x > \alpha^t_{1:\ell(x)}.$$

On-sequence, i.e. for $x = \alpha_{1:n}$, ν^t is somewhere in between 0 and $2^{-\ell(x)}$. Since sequence $\alpha := \lim_t \alpha^t$ is λ.M.L.-random it contains 01 infinitely often, actually $\alpha_n\alpha_{n+1} = 01$ for a non-vanishing fraction of n. In the following we fix such an n. For $t \geq n$ we get

$$\nu^t(\alpha_{<n}) = \nu^t(\alpha_{<n}0) + \nu^t(\underbrace{\alpha_{<n}1}_{>\alpha_{1:n}\geq\alpha^t_{1:n},\ \text{since}\ \alpha_n=0}) = \nu^t(\alpha_{<n}0) = \nu^t(\alpha_{1:n}) \Rightarrow \nu(\alpha_{<n}) = \nu(\alpha_{1:n}).$$

This ensures $\nu(\alpha_n|\alpha_{<n}) = 1 \neq \frac{1}{2} = \lambda_n$. For $t > n$ large enough such that $\alpha^t_{1:n+1} = \alpha_{1:n+1}$ we get:

$$\nu^t(\alpha_{1:n}) = \nu^t(\alpha^t_{1:n}) \geq \nu^t(\underbrace{\alpha^t_{1:n}0}_{<\alpha^t_{1:n+1},\ \text{since}\ \alpha_{n+1}=1}) = 2^{-n-1} \Rightarrow \nu(\alpha_{1:n}) \geq 2^{-n-1}.$$

This ensures $\nu(\alpha_{1:n}) \geq 2^{-n-1} \geq \frac{1}{2}M(\alpha_{1:n})$ by (5). Let M be any universal semimeasure and $0 < \gamma < \frac{1}{5}$. Then $M'(x) := (1 - \gamma)\nu(x) + \gamma M(x)\,\forall x$ is also a universal semimeasure with

$$M'(\alpha_n|\alpha_{<n}) \;=\; \frac{(1-\gamma)\nu(\alpha_{1:n}) + \gamma M(\alpha_{1:n})}{(1-\gamma)\nu(\alpha_{<n}) + \gamma M(\alpha_{<n})} \;\overset{\substack{M(\alpha_{<n}) \leq 2^{-n+1} \text{ and } M(\alpha_{1:n}) \geq 0 \\ \downarrow}}{\geq}\; \frac{(1-\gamma)\nu(\alpha_{1:n})}{(1-\gamma)\nu(\alpha_{<n}) + \gamma 2^{-n+1}}$$

$$\underset{\substack{\uparrow \\ \nu(\alpha_{<n})=\nu(\alpha_{1:n})}}{=} \quad \frac{1-\gamma}{1-\gamma+\gamma 2^{-n+1}/\nu(\alpha_{1:n})} \quad \underset{\substack{\uparrow \\ \nu(\alpha_{1:n})\geq 2^{-n-1}}}{\geq} \quad \frac{1-\gamma}{1+3\gamma} \quad > \quad \frac{1}{2}.$$

For instance for $\gamma = \frac{1}{9}$ we have $M'(\alpha_n|\alpha_{<n}) \geq \frac{2}{3} \neq \frac{1}{2} = \lambda(\alpha_n|\alpha_{<n})$ for a non-vanishing fraction of n's. Note that the contamination of M with ν must be sufficiently large (γ sufficiently small), while an advantage of the non-constructive proof is that an arbitrarily small contamination sufficed. □

A converse of Theorem 6 can also be shown:

Theorem 10 (*Convergence on Non-random Sequences*). *For every universal semimeasure M there exist computable measures μ and non-μ.M.L.-random sequences α for which $M(\alpha_n|\alpha_{<n})/\mu(\alpha_n|\alpha_{<n}) \to 1$.*

5. Convergence in Martin-Löf sense

In this section we give a positive answer to the question of predictive M.L.-convergence to μ. We consider general finite alphabet $\mathcal{X}$.

Theorem 11 (*Universal Predictor for M.L.-Random Sequences*). *There exists an enumerable semimeasure W such that for every computable measure μ and every μ.M.L.-random sequence ω, the predictions converge to each other:*

$$W(a|\omega_{<t}) \overset{t\to\infty}{\longrightarrow} \mu(a|\omega_{<t}) \quad \textit{for all} \quad a \in \mathcal{X} \quad \textit{if} \quad d_\mu(\omega) < \infty.$$

The semimeasure W we will construct is not universal in the sense of dominating all enumerable semimeasures, unlike M. Normalizing W shows that there is also a measure whose predictions converge to μ, but this measure is not enumerable, only approximable. For proving Theorem 11 we first define an intermediate measure D as a mixture over all computable measures, which is not even approximable. Based on Lemmas 4, 12 and 13, Proposition 14 shows that D M.L.-converges to μ. We then define the concept of quasimeasures in Definition 15 and an enumerable semimeasure W as a mixture over all enumerable quasimeasures. Proposition 18 shows that W M.L.-converges to D. Theorem 11 immediately follows from Propositions 14 and 18.

Lemma 12 (*Hellinger Chain*). *Let $h(p,q) := \sum_{i=1}^N (\sqrt{p_i} - \sqrt{q_i})^2$ be the Hellinger distance between $p = (p_i)_{i=1}^N \in \mathbb{R}_+^N$ and $q = (q_i)_{i=1}^N \in \mathbb{R}_+^N$. Then*

$$\begin{array}{llcl} \text{(i)} \quad \textit{for } p,q,r \in \mathbb{R}_+^N & h(p,q) & \leq & (1+\beta)\,h(p,r) + (1+\beta^{-1})\,h(r,q), \ \textit{any } \beta > 0 \\ \text{(ii)} \quad \textit{for } p^1,\dots,p^m \in \mathbb{R}_+^N & h(p^1,p^m) & \leq & 3\sum_{k=2}^{m} k^2\,h(p^{k-1},p^k). \end{array}$$

Proof. (i) For any $x,y,z \in \mathbb{R}$ and $\beta > 0$, squaring the triangle inequality $|x-y| \leq |x-z| + |z-y|$ and chaining it with the binomial $2|x-z||z-y| \leq \beta(x-z)^2 + \beta^{-1}(z-y)^2$ shows $(x-y)^2 \leq (1+\beta)(x-z)^2 + (1+\beta^{-1})(z-y)^2$. (i) follows for $x = \sqrt{p_i}$, $y = \sqrt{q_i}$, and $z = \sqrt{r_i}$ and summation over i.

(ii) Applying (i) for the triples (p^k, p^{k+1}, p^m) for and in order of $k = 1, 2, \dots, m-2$ with $\beta = \beta_k$ gives

$$h(p^1,p^m) \leq \sum_{k=2}^{m} \left[\prod_{j=1}^{k-2} (1+\beta_j^{-1})\right] \cdot (1+\beta_{k-1}) \cdot h(p^{k-1},p^k).$$

For $\beta_k = k(k+1)$ we have $\ln \prod_{j=1}^{k-2}(1+\beta_j^{-1}) \leq \sum_{j=1}^{\infty} \ln(1+\beta_j^{-1}) \leq \sum_{j=1}^{\infty} \beta_j^{-1} = 1$ and $1 + \beta_{k-1} \leq k^2$, which completes the proof. The choice $\beta_k = 2^{K(k)}$ would lead to a bound with $1 + 2^{K(k)}$ instead of k^2. □

We need a way to convert expected bounds to bounds on individual M.L. random sequences, sort of a converse of "M.L. implies w.p.1". Consider for instance the Hellinger sum $H(\omega) := \sum_{t=1}^{\infty} h_t(\mu,\rho)/\ln w^{-1}$ between two computable measures $\rho \geq w \cdot \mu$. Then H is an enumerable function and Lemma 4 implies $\mathbf{E}[H] \leq 1$, hence H is an integral μ-test. H can be increased to an enumerable μ-supermartingale $\bar{H}$. The universal μ-supermartingale M/μ multiplicatively dominates all enumerable supermartingales (and hence $\bar{H}$). Since $M/\mu \leq 2^{d_\mu(\omega)}$, this implies the desired bound $H(\omega)\overset{\times}{\leq}2^{d_\mu(\omega)}$ for individual ω. We give a self-contained direct proof, explicating all important constants.

Lemma 13 (*Expected to Individual Bound*). *Let $F(\omega)\geq 0$ be an enumerable function and μ be an enumerable measure and $\varepsilon>0$ be co-enumerable. Then:*

$$\textit{If}\quad \mathbf{E}_\mu[F]\leq\varepsilon\quad\textit{then}\quad F(\omega)\;\overset{\times}{\leq}\;\varepsilon\cdot 2^{K(\mu,F,{}^1\!/\varepsilon)+d_\mu(\omega)}\quad\forall\omega$$

where $d_\mu(\omega)$ is the μ-randomness deficiency of ω and $K(\mu,F,{}^1\!/\varepsilon)$ is the length of the shortest program for μ, F, and ${}^1\!/\varepsilon$.

Lemma 13 roughly says that for μ, F, and $\varepsilon\overset{\times}{=}\mathbf{E}_\mu[F]$ with short program ($K(\mu,F,{}^1\!/\varepsilon)=O(1)$) and μ-random ω ($d_\mu(\omega)=O(1)$) we have $F(\omega)\overset{\times}{\leq}\mathbf{E}_\mu[F]$.

Proof. Let $F(\omega)=\lim_{n\to\infty}F_n(\omega)=\sup_n F_n(\omega)$ be enumerated by an increasing sequence of computable functions $F_n(\omega)$. $F_n(\omega)$ can be chosen to depend on $\omega_{1:n}$ only, i.e. $F_n(\omega)=F_n(\omega_{1:n})$ is independent of $\omega_{n+1:\infty}$. Let $\varepsilon_n\searrow\varepsilon$ co-enumerate ε. We define

$$\bar\mu_n(\omega_{1:k})\;:=\;\varepsilon_n^{-1}\sum_{\omega_{k+1:n}\in\mathcal{X}^{n-k}}\mu(\omega_{1:n})F_n(\omega_{1:n})\;\;\text{for}\;\;k\leq n,\qquad\text{and}\qquad\bar\mu_n(\omega_{1:k})=0\;\;\text{for}\;\;k>n.$$

$\bar\mu_n$ is a computable semimeasure for each n (due to $\mathbf{E}_\mu[F_n]\leq\varepsilon$) and increasing in n, since

$$\begin{aligned}&\bar\mu_n(\omega_{1:k})\geq\;\;0=\bar\mu_{n-1}(\omega_{1:k})\quad\text{for}\quad k\geq n\quad\text{and}\\ &\bar\mu_n(\omega_{<n})\;\underset{\substack{\uparrow\\ F_n\geq F_{n-1}}}{\geq}\sum_{\omega_n\in\mathcal{X}}\varepsilon_n^{-1}\mu(\omega_{1:n})F_{n-1}(\omega_{<n})\;\underset{\substack{\uparrow\\ \mu\text{ measure}}}{=}\;\varepsilon_n^{-1}\mu(\omega_{<n})F_{n-1}(\omega_{<n})\;\underset{\substack{\uparrow\\ \varepsilon_n\leq\varepsilon_{n-1}}}{\geq}\;\bar\mu_{n-1}(\omega_{<n})\end{aligned}$$

and similarly for $k<n-1$. Hence $\bar\mu:=\bar\mu_\infty$ is an enumerable semimeasure (indeed $\bar\mu$ is proportional to a measure). From dominance (2) we get

$$M(\omega_{1:n})\;\overset{\times}{\geq}\;2^{-K(\bar\mu)}\bar\mu(\omega_{1:n})\geq 2^{-K(\bar\mu)}\bar\mu_n(\omega_{1:n})=2^{-K(\bar\mu)}\varepsilon_n^{-1}\mu(\omega_{1:n})F_n(\omega_{1:n}).\tag{7}$$

In order to enumerate $\bar\mu$, we need to enumerate μ, F, and ε^{-1}, hence $K(\bar\mu)\overset{+}{\leq}K(\mu,F,{}^1\!/\varepsilon)$, so we get

$$F_n(\omega)\equiv F_n(\omega_{1:n})\;\overset{\times}{\leq}\;\varepsilon_n\cdot 2^{K(\mu,F,{}^1\!/\varepsilon)}\cdot\tfrac{M(\omega_{1:n})}{\mu(\omega_{1:n})}\leq\varepsilon_n\cdot 2^{K(\mu,F,{}^1\!/\varepsilon)+d_\mu(\omega)}.$$

Taking the limit $F_n\nearrow F$ and $\varepsilon_n\searrow\varepsilon$ completes the proof. □

Let $\mathcal{M}=\{\nu_1,\nu_2,\ldots\}$ be an enumeration of all enumerable semimeasures, $J_k:=\{i\leq k:\nu_i\text{ is measure}\}$, and $\delta_k(x):=\sum_{i\in J_k}\varepsilon_i\nu_i(x)$. The weights ε_i need to be computable and exponentially decreasing in i and $\sum_{i=1}^\infty\varepsilon_i\leq 1$. We choose $\varepsilon_i=i^{-6}2^{-i}$. Note the subtle and important fact that although the definition of J_k is non-constructive, as a finite set of finite objects, J_k is decidable (the program is unknowable for large k). Hence, δ_k is computable, since enumerable measures are computable.

$$D(x)=\delta_\infty(x)=\sum_{i\in J_\infty}\varepsilon_i\nu_i(x)=\text{mixture of all computable measures.}$$

In contrast to J_k and δ_k, the set J_∞ and hence D are neither enumerable nor co-enumerable. We also define the measures $\hat\delta_k(x):=\delta_k(x)/\delta_k(\epsilon)$ and $\hat D(x):=D(x)/D(\epsilon)$. The following proposition implies predictive convergence of D to μ on μ-random sequences.

Proposition 14 (*Convergence of Incomputable Measure $\hat D$*). *Let μ be a computable measure with index k_0, i.e. $\mu=\nu_{k_0}$. Then for the incomputable measure $\hat D$ and the computable but non-constructive measures $\hat\delta_{k_0}$ defined above, the following holds:*

$$\begin{array}{lllll}\text{(i)} & \sum_{t=1}^\infty h_t(\hat\delta_{k_0},\mu) & \overset{+}{\leq} & 2\ln 2\cdot d_\mu(\omega)+3k_0\\ \text{(ii)} & \sum_{t=1}^\infty h_t(\hat\delta_{k_0},\hat D) & \overset{\times}{\leq} & k_0^7 2^{k_0+d_\mu(\omega)}.\end{array}$$

Combining (i) and (ii), using Lemma 12(i), we get $\sum_{t=1}^{\infty} h_t(\mu,\hat{D}) \leq c_\omega f(k_0) < \infty$ for μ-random ω, which implies $D(b|\omega_{<t}) \equiv \hat{D}(b|\omega_{<t}) \to \mu(b|\omega_{<t})$. We do not know whether on-sequence convergence of the ratio holds. Similar bounds hold for $\hat{\delta}_{k_1}$ instead $\hat{\delta}_{k_0}$, $k_1 \geq k_0$. The principle proof idea is to convert the expected bounds of Lemma 4 to individual bounds, using Lemma 13. The problem is that $\hat{D}$ is not computable, which we circumvent by joining with Lemma 12, bounds on $\sum_t h_t(\hat{\delta}_{k-1},\hat{\delta}_k)$ for $k = k_0, k_0+1, \ldots$.

Proof. (i) Let $H(\omega) := \sum_{t=1}^{\infty} h_t(\hat{\delta}_{k_0},\mu)$. μ and $\hat{\delta}_{k_0}$ are measures with $\hat{\delta}_{k_0} \geq \delta_{k_0} \geq \varepsilon_{k_0}\mu$, since $\delta_k(\epsilon) \leq 1$, $\mu = \nu_{k_0}$ and $k_0 \in J_{k_0}$. Hence, Lemma 4 applies and shows $\mathbf{E}_\mu[\exp(\frac{1}{2}H)] \leq \varepsilon_{k_0}^{-1/2}$. H is well defined and enumerable for $d_\mu(\omega) < \infty$, since $d_\mu(\omega) < \infty$ implies $\mu(\omega_{1:t}) \neq 0$ implies $\hat{\delta}_{k_0}(\omega_{1:t}) \neq 0$. So $\mu(b|\omega_{1:t})$ and $\hat{\delta}_{k_0}(b|\omega_{1:t})$ are well defined and computable (given J_{k_0}). Hence $h_t(\hat{\delta}_{k_0},\mu)$ is computable, hence $H(\omega)$ is enumerable. Lemma 13 then implies $\exp(\frac{1}{2}H(\omega)) \overset{\times}{\leq} \varepsilon_{k_0}^{-1/2} \cdot 2^{K(\mu,H,\sqrt{\varepsilon}_{k_0})+d_\mu(\omega)}$. We bound

$$K(\mu,H,\sqrt{\varepsilon}_{k_0}) \overset{+}{\leq} K(H|\mu,k_0)+K(k_0) \overset{+}{\leq} K(J_{k_0}|k_0)+K(k_0) \overset{+}{\leq} k_0+2\log k_0.$$

The first inequality holds, since k_0 is the index and hence a description of μ, and $\varepsilon_{()}$ is a simple computable function. H can be computed from μ, k_0 and J_{k_0}, which implies the second inequality. The last inequality follows from $K(k_0) \overset{+}{\leq} 2\log k_0$ and the fact that for each $i \leq k_0$ one bit suffices to specify (non)membership to J_{k_0}, i.e. $K(J_{k_0}|k_0) \overset{+}{\leq} k_0$. Putting everything together we get

$$H(\omega) \overset{+}{\leq} \ln\varepsilon_{k_0}^{-1} + [k_0+2\log k_0+d_\mu(\omega)]2\ln 2 \overset{+}{\leq} (2\ln 2)d_\mu(\omega)+3k_0.$$

(ii) Let $H^k(\omega) := \sum_{t=1}^{\infty} h_t(\hat{\delta}_k,\hat{\delta}_{k-1})$ and $k > k_0$. $\delta_{k-1} \leq \delta_k$ implies

$$\frac{\hat{\delta}_{k-1}(x)}{\hat{\delta}_k(x)} \leq \frac{\delta_k(\epsilon)}{\delta_{k-1}(\epsilon)} \leq \frac{\delta_{k-1}(\epsilon)+\varepsilon_k}{\delta_{k-1}(\epsilon)} = 1+\frac{\varepsilon_k}{\delta_{k-1}(\epsilon)} \leq 1+\frac{\varepsilon_k}{\varepsilon_O},$$

where $O := \min\{i \in J_{k-1}\} = O(1)$. Note that $J_{k-1} \ni k_0$ is not empty. Since $\hat{\delta}_{k-1}$ and $\hat{\delta}_k$ are measures, Lemma 4 applies and shows $\mathbf{E}_{\hat{\delta}_{k-1}}[H^k] \leq \ln(1+\frac{\varepsilon_k}{\varepsilon_O}) \leq \frac{\varepsilon_k}{\varepsilon_O}$. Exploiting $\varepsilon_{k_0}\mu \leq \hat{\delta}_{k-1}$, this implies $\mathbf{E}_\mu[H^k] \leq \frac{\varepsilon_k}{\varepsilon_O\varepsilon_{k_0}}$. Lemma 13 then implies $H^k(\omega) \overset{\times}{\leq} \frac{\varepsilon_k}{\varepsilon_O\varepsilon_{k_0}} \cdot 2^{K(\mu,H^k,\varepsilon_O\varepsilon_{k_0}/\varepsilon_k)+d_\mu(\omega)}$. Similarly as in (i) we can bound

$$K(\mu,H^k,\varepsilon_{k_0}/\varepsilon_O\varepsilon_k) \overset{+}{\leq} K(J_k|k)+K(k)+K(k_0) \overset{+}{\leq} k+2\log k+2\log k_0, \quad \text{hence}$$
$$H^k(\omega) \overset{\times}{\leq} \frac{\varepsilon_k}{\varepsilon_O\varepsilon_{k_0}} \cdot k_0^2k^22^kc_\omega \overset{\times}{=} k_0^82^{k_0}k^{-4}c_\omega, \quad \text{where} \quad c_\omega := 2^{d_\mu(\omega)}.$$

Chaining this bound via Lemma 12(ii) we get for $k_1 > k_0$:

$$\begin{aligned}\sum_{t=1}^{n} h_t(\hat{\delta}_{k_0},\hat{\delta}_{k_1}) &\leq \sum_{t=1}^{n} 3 \sum_{k=k_0+1}^{k_1} (k-k_0+1)^2 h_t(\hat{\delta}_{k-1},\hat{\delta}_k)\\ &\leq 3\sum_{k=k_0+1}^{k_1} k^2H^k(\omega) \overset{\times}{\leq} 3k_0^82^{k_0}c_\omega \sum_{k=k_0+1}^{k_1} k^{-2} \leq 3k_0^72^{k_0}c_\omega.\end{aligned}$$

If we now take $k_1 \to \infty$ we get $\sum_{t=1}^{n} h_t(\hat{\delta}_{k_0},\hat{D}) \overset{\times}{\leq} 3k_0^72^{k_0+d_\mu(\omega)}$. Finally let $n \to \infty$. □

The main properties allowing for proving $\hat{D} \to \mu$ were that $\hat{D}$ is a measure with approximations $\hat{\delta}_k$, which are computable in a certain sense. $\hat{D}$ is a mixture over all enumerable/computable measures and hence incomputable.

6. M.L.-converging enumerable semimeasure W

The next step is to enlarge the class of computable measures to an enumerable class of semimeasures, which are still sufficiently close to measures in order not to spoil the convergence result. For convergence w.p.1. we could include *all* semimeasures (Theorem 3). M.L.-convergence seems to require a more restricted class. Included non-measures need to be zero on long strings. We define quasimeasures as nearly normalized measures on $\mathcal{X}^{\leq n}$.

Definition 15 (*Quasimeasures*). $\tilde\nu:\mathcal{X}^*\to\mathbb{R}_+$ is called a quasimeasure *iff* $\tilde\nu$ is a measure or: $\sum_{a\in\mathcal{X}}\tilde\nu(xa)=\tilde\nu(x)$ for $\ell(x)<n$ and $\tilde\nu(x)=0$ for $\ell(x)>n$ and $1-\frac{1}{n}<\tilde\nu(\epsilon)\leq 1$, for some $n\in\mathbb{N}$.

Lemma 16 (*Quasimeasures*). (i) *A quasimeasure is either a semimeasure which is zero on long strings -or- a measure.* (ii) *The set of enumerable quasimeasures is enumerable and contains all computable measures.*

For enumerability it is important to include the measures in the definition of quasimeasures. One way of enumeration would be to enumerate all enumerable partial functions f and convert them to quasimeasures. Since we need a correspondence to semimeasures, we convert a semimeasure ν directly to a maximal quasimeasure $\tilde\nu\leq\nu$.

Proof & construction. (i) Obvious from Definition 15.

(ii) Let ν be an enumerable semimeasure enumerated by $\nu^t\nearrow\nu$. Consider $m\equiv m^t:=\max\{n\leq t:\sum_{x_{1:n}}\nu^t(x_{1:n})>1-\frac{1}{n}\}$. m^t is finite and monotone increasing in t. We define the quasimeasure

$$\rho^t(x_{1:n}):=\sum_{x_{n+1:m}\in\mathcal{X}^{m-n}}\nu^t(x_{1:m})\quad\text{for}\quad n\leq m\quad\text{and}\quad\rho^t(x_{1:n})=0\quad\text{for}\quad n>m.$$

We define an increasing sequence in t of quasimeasures $\tilde\nu^t\leq\nu^t$ for $t=1,2,\ldots$ recursively starting with $\tilde\nu^0:=0$ as follows:

If $\rho^t(x_{1:n})\geq\tilde\nu^{t-1}(x_{1:n})\ \forall x_{1:n}\forall n\leq m^t$ (and hence $\forall x$), then $\tilde\nu^t:=\rho^t$, else $\tilde\nu^t:=\tilde\nu^{t-1}$.

$\tilde\nu:=\lim_{t\to\infty}\tilde\nu^t$ is an enumerable quasimeasure. Note that $m^\infty=\infty$ iff ν is a measure. One can easily verify that $\tilde\nu\leq\nu$ and $\tilde\nu\equiv\nu$ iff ν is a quasimeasure. This implies that if $\nu_1,\nu_2,\ldots$ is an enumeration of all enumerable semimeasures, then $\tilde\nu_1,\tilde\nu_2,\ldots$ is an enumeration of all enumerable quasimeasures. □

Let $\tilde\nu_1,\tilde\nu_2,\ldots$ be the enumeration of all enumerable quasimeasures constructed in the proof of Lemma 16, based on the enumeration of all enumerable semimeasures $\nu_1,\nu_2,\ldots$ with the property that $\tilde\nu_i\leq\nu_i$ and equality holds if ν_i is a (quasi)measure. We define the enumerable semimeasure

$$W(x):=\sum_{i=1}^\infty\varepsilon_i\tilde\nu_i(x),\qquad\text{and note that}\quad D(x)=\sum_{i\in J}\varepsilon_i\tilde\nu_i(x)\ \text{ with }\ J:=\{i:\tilde\nu_i\text{ is measure}\}$$

with $\varepsilon_i=i^{-6}2^{-i}$ as before. To show $W\to D$ we need the following lemma.

Lemma 17 (*Hellinger Continuity*). *For $h_x(\mu,\rho):=\sum_{a\in\mathcal{X}}(\sqrt{\mu(a|x)}-\sqrt{\rho(a|x)})^2$, where $\rho(y)=\mu(y)+\nu(y)$ $\forall y\in\mathcal{X}^*$ and μ and ν are semimeasures, it holds:*

(i) $h_x(\mu,\rho)\leq\frac{\nu(x)}{\mu(x)}$.

(ii) $h_x(\mu,\rho)\leq\frac{1}{4}\varepsilon^2$ *if* $\nu(x)\leq\varepsilon\cdot\mu(x)$ *and* $\nu(xb)\leq\varepsilon\cdot\mu(xb)\ \forall b\in\mathcal{X}$.

(ii) Since the Hellinger distance is locally quadratic, $h_x(\mu,\rho)$ scales quadratic in the deviation of predictor ρ from μ. (i) Closeness of $\rho(x)$ to $\mu(x)$ only, does not imply closeness of the predictions, hence only a bound linear in the deviation is possible.

Proof. (i) We identify $\mathcal{X}\cong\{1,\ldots,N\}$ and define $y_i=\mu(xi)$, $z_i=\nu(xi)$, $y=\mu(x)$, and $z=\nu(x)$. We extend $(y_i)_{i=1}^N$ to a probability by defining $y_0=y-\sum_{i=1}^N y_i\geq 0$ and set $z_0=0$. Also $\varepsilon':=z/y$. Exploiting $\sum_{i=0}^N y_i=y$ and $\sum_{i=0}^N z_i\leq z$ and $z\leq\varepsilon y$ and $y_i,z_i,y,z\geq 0$ we get

$$h_x(\mu,\mu+\nu)\equiv\sum_{i=1}^N\left(\sqrt{\frac{y_i}{y}}-\sqrt{\frac{y_i+z_i}{y+z}}\right)^2\leq\sum_{i=0}^N\left(\sqrt{\frac{y_i}{y}}-\sqrt{\frac{y_i+z_i}{y+z}}\right)^2$$

$$=\sum_{i=0}^N\left(\frac{y_i}{y}+\frac{y_i+z_i}{y+z}-2\sqrt{\frac{y_i(y_i+z_i)}{y(y+z)}}\right)\leq 2-2\sum_{i=0}^N\frac{y_i}{\sqrt{y(y+z)}}=2-\frac{2}{\sqrt{1+\varepsilon'}}\leq\varepsilon'.$$

(ii) With the notation from (i), additionally exploiting $z_i \leq \varepsilon y_i$ we get

$$\sqrt{\frac{y_i+z_i}{y+z}} - \sqrt{\frac{y_i}{y}} \leq \frac{\sqrt{y_i+z_i} - \sqrt{y_i}}{\sqrt{y}} \leq \frac{\sqrt{y_i(1+\varepsilon)} - \sqrt{y_i}}{\sqrt{y}} \leq \frac{\varepsilon}{2}\sqrt{\frac{y_i}{y}} \qquad \text{and}$$

$$\sqrt{\frac{y_i}{y}} - \sqrt{\frac{y_i+z_i}{y+z}} = \frac{\sqrt{y_i(1+\varepsilon')} - \sqrt{y_i+z_i}}{\sqrt{y(1+\varepsilon')}} \leq \frac{\sqrt{y_i(1+\varepsilon')} - \sqrt{y_i}}{\sqrt{y(1+\varepsilon')}} \leq \frac{\varepsilon'}{2}\sqrt{\frac{y_i}{y}}.$$

Exploiting $\varepsilon' \leq \varepsilon$, taking the square and summing over i proves (ii). □

Proposition 18 (*Convergence of Enumerable W to Incomputable D*). *For every computable measure μ and for ω being μ-random, the following holds for $t \to \infty$:*

$$\text{(i)}\ \frac{W(\omega_{1:t})}{D(\omega_{1:t})} \to 1, \qquad \text{(ii)}\ \frac{W(\omega_t|\omega_{<t})}{D(\omega_t|\omega_{<t})} \to 1, \qquad \text{(iii)}\ W(a|\omega_{<t}) \to D(a|\omega_{<t})\ \forall a \in \mathcal{X}.$$

The intuitive reason for the convergence is that the additional contributions of non-measures to W absent in D are zero for long sequences.

Proof. (i)

$$D(x) \leq W(x) = D(x) + \sum_{i \notin J} \varepsilon_i \tilde{\nu}_i(x) \leq D(x) + \sum_{i=k_x}^{\infty} \varepsilon_i \tilde{\nu}_i(x), \tag{8}$$

where $k_x := \min_i\{i \notin J : \tilde{\nu}_i(x) \neq 0\}$. For $i \notin J$, $\tilde{\nu}_i$ is not a measure. Hence $\tilde{\nu}_i(x) = 0$ for sufficiently long x. This implies $k_x \to \infty$ for $\ell(x) \to \infty$, hence $W(x) \to D(x)\ \forall x$. To get convergence in ratio we have to assume that $x = \omega_{1:n}$ with ω being μ-random, i.e. $c_\omega := \sup_n \frac{M(\omega_{1:n})}{\mu(\omega_{1:n})} = 2^{d_\mu(\omega)} < \infty$.

$$\Rightarrow\ \tilde{\nu}_i(x) \leq \nu_i(x) \leq \frac{1}{w_{\nu_i}} M(x) \leq \frac{c_\omega}{w_{\nu_i}} \mu(x) \leq \frac{c_\omega}{w_{\nu_i}\varepsilon_{k_0}} D(x).$$

The last inequality holds, since μ is a computable measure of index k_0, i.e. $\mu = \nu_{k_0} = \tilde{\nu}_{k_0}$. Inserting $1/w_{\nu_i} \leq c' \cdot i^2$ for some $c = O(1)$ and ε_i we get $\varepsilon_i \tilde{\nu}_i(x) \leq \frac{c' c_\omega}{\varepsilon_{k_0}} i^{-4} 2^{-i} D(x)$, which implies $\sum_{i=k_x}^{\infty} \varepsilon_i \tilde{\nu}_i(x) \leq \varepsilon'_x D(x)$ with

$$\varepsilon'_x := \frac{c' c_\omega}{\varepsilon_{k_0}} \sum_{i=k_x}^{\infty} i^{-4} 2^{-i} \leq \frac{2c' c_\omega}{\varepsilon_{k_0}} k_x^{-4} 2^{-k_x} \to 0 \quad \text{for} \quad \ell(x) \to \infty.$$

Inserting this into (8) we get

$$1 \leq \frac{W(x)}{D(x)} \leq 1 + \varepsilon'_x \overset{\ell(x)\to\infty}{\longrightarrow} 1 \quad \text{for } \mu\text{-random } x.$$

(ii) Obvious from (i) by taking a double ratio.

(iii) Since D and $W - D$ are semimeasures and $\frac{W-D}{W} \leq \varepsilon'_x$ by (i), Lemma 17(i) implies $h_x(D, W) \leq \varepsilon'_x$. Since $\varepsilon'_x \to 0$ for μ-random x, this shows (iii). $|W(a|x) - D(a|x)| \leq \varepsilon'_x$ can also be shown. □

Speed of convergence.

The main convergence Theorem 11 now immediately follows from Propositions 14 and 18. We briefly remark on the convergence rate. For M, Lemma 4 shows that $\mathbf{E}[\sum_t h_t(M, \mu)] \leq \ln w_{k_0}^{-1} \overset{\times}{=} \ln k_0$ is logarithmic in the index k_0 of μ, but $\mathbf{E}[\sum_t h_t(X, \mu)] \leq \ln \varepsilon_{k_0} \overset{\times}{=} k_0$ is linear in k_0 for $X = [W, D, \delta_{k_0}]$. The individual bounds for $\sum_t h_t(\hat{\delta}_{k_0}, \mu)$ and $\sum_t h_t(\hat{\delta}_{k_0}, \hat{D})$ in Proposition 14 are linear and exponential in k_0, respectively. For $W \overset{M.L.}{\longrightarrow} D$ we could not establish any convergence speed.

Finally we show that W does not dominate all enumerable semimeasures, as the definition of W suggests. We summarize all computability, measure, and dominance properties of M, D, $\hat{D}$, and W in the following theorem:

Theorem 19 (*Properties of M, W, D, and $\hat{D}$*).
(i) *M is an enumerable semimeasure, which dominates all enumerable semimeasures. M is not computable and not a measure.*
(ii) *$\hat{D}$ is a measure, D is proportional to a measure, both dominating all enumerable quasimeasures. D and $\hat{D}$ are not computable and do not dominate all enumerable semimeasures.*
(iii) *W is an enumerable semimeasure, which dominates all enumerable quasimeasures. W is not itself a quasimeasure, is not computable, and does not dominate all enumerable semimeasures.*

We conjecture that D and $\hat{D}$ are not even approximable (limit-computable), but lie somewhere higher in the arithmetic hierarchy. Since W can be normalized to an approximable measure M.L.-converging to μ, and D was only an intermediate quantity, the question of approximability of D seems not too interesting.

Proof. (i) First sentence: Holds by definition. That such an M exists follows from the enumerability of all enumerable semimeasures [17,8]. Second sentence: If M were a measure it would be computable, contradicting [3, Thm. 4(iii)] (see below).

(ii) First sentence: Follows from the definition of D and $\hat{D}$ and the fact that quasimeasures are zero on long strings: $\frac{D}{\nu} \geq \varepsilon_\nu > 0$ if ν is a computable measure. If ν is a "proper" quasimeasure, then $\min_{x\in\mathcal{X}^*} \frac{D(x)}{\nu(x)} = \min_{x:\ell(x)\leq m_\nu} \frac{D(x)}{\nu(x)} > 0$, since $\nu(x) = 0$ for $\ell(x) > m_\nu < \infty$, and $D(x) > 0\,\forall x$. Second sentence: It is well known that there is no computable semimeasure dominating all computable measures (see e.g. [3, Thm. 4]), which shows that D, $\hat{D}$ and W cannot be computable. We now show that D and W do not dominate the enumerable semimeasure M by extending this argument. Let ν be a nowhere[2] zero computable semimeasure. We define a computable sequence α as follows by induction: Given $\alpha_{<n}$, choose some α_n in a computable way (by computing ν to sufficient accuracy) such that $\nu(\alpha_n|\alpha_{<n}) < |\mathcal{X}|^{-1}(1+\frac{1}{n^2})$. Such an α_n exists, since ν is a semimeasure. We then define the computable deterministic measure $\bar{\nu}$ concentrated on α, i.e. $\bar{\nu}(\alpha_{1:n}) = 1\ \forall n$ and $\bar{\nu}(x) = 0$ for all x which are not prefixes of α. By the chain rule we get $\nu(\alpha_{1:n}) \leq \frac{\sinh\pi}{\pi}|\mathcal{X}|^{-n} \leq 4|\mathcal{X}|^{-n}\bar{\nu}(\alpha_{1:n})$. This shows that no computable semimeasure ν can dominate all computable measures, since $\bar{\nu}$ is not dominated. We use this construction for $\nu = \delta_k$:

$$\begin{aligned}\sum_{i=1}^{k}\varepsilon_i\tilde{\nu}_i(\alpha_{1:n}) \;&\overset{\substack{\text{for sufficiently large } n=n_k\\ \downarrow}}{=}\; \delta_k(\alpha_{1:n}) \leq 4|\mathcal{X}|^{-n}\bar{\delta}_k(\alpha_{1:n}) \;\overset{\substack{M\overset{\times}{\geq}2^{-K(\nu)}\nu\\ \downarrow}}{\overset{\times}{\leq}}\; |\mathcal{X}|^{-n}2^{K(\bar{\delta}_k)}M(\alpha_{1:n})\\ &\underset{\substack{\uparrow\\ K(\bar{\delta}_k)\overset{+}{\leq}K(\delta_k)\overset{+}{\leq}k+2\log k}}{\overset{\times}{\leq}}\; |\mathcal{X}|^{-n}k^2 2^k M(\alpha_{1:n}) \;\underset{\substack{\uparrow\\ \text{for } n\geq\frac{2}{\log|\mathcal{X}|}k}}{\leq}\; k^2 2^{-k}M(\alpha_{1:n}).\end{aligned} \tag{9}$$

For all x we have

$$D(x) - \delta_k(x) \leq \sum_{i=k+1}^{\infty}\varepsilon_i\tilde{\nu}_i(x) = \sum_{i=k+1}^{\infty} i^{-6}2^{-i}\tilde{\nu}_i(x) \leq 2^{-k}\sum_{i=k+1}^{\infty} i^{-6}\nu_i(x) \overset{\times}{\leq} 2^{-k}M(x).$$

Summing both bounds we get $D(\alpha_{1:n_k}) \leq W(\alpha_{1:n_k}) \overset{\times}{\leq} (k^2+1)2^{-k}M(\alpha_{1:n_k})$, which shows that D, $\hat{D}$ and W do not dominate the enumerable semimeasure M.

Remark: Note that the constructed sequence(s) α depends on the choice of k, so we should write more precisely $\alpha = \alpha^k$. For D (but not for W) we can choose $k = \frac{n}{2}\log|\mathcal{X}|$ in (9) (satisfying $n \geq \frac{2}{\log|\mathcal{X}|}k$), leading to $D(\alpha^n_{1:n}) \overset{\times}{\leq} n^2|\mathcal{X}|^{-n/2}M(\alpha^n_{1:n})$. It is easy to generalize (9) to $\forall x_{<t}\exists\alpha_{t:n} : \delta_k(x_{<t}\alpha_{t:n}) \overset{\times}{\leq} |\mathcal{X}|^{t-n}k^2 2^k M(x_{<t}\alpha_{t:n})$, where t is a simple function of k. Choosing $t = k^2+1$ and $n = (k+1)^2$ and joining the results for $k = 1, 2, \ldots$ and $x_{<t} := \alpha_{<t}$ we get $D(\alpha_{1:n}) \overset{\times}{\leq} n2^{-\sqrt{n}}M(\alpha_{1:n})\ \forall n$ for the single sequence α. This implies that (but is stronger than) α is not random w.r.t. to any computable measure $\tilde{\nu}$. Such α are sometimes called absolutely non-stochastic.

[2] M, W, $\hat{D}$, D, and δ_k for $k \geq O(1)$ are nowhere zero. Alternatively one can verify that all relevant assertions remain valid if ν is somewhere zero.

(iii) First sentence: Enumerability is immediate from the definition, given the enumerability of all enumerable quasimeasures. Second sentence: Since quasimeasures drop out in the mixture defining W for long x, W cannot be a measure. Since $W(x) \neq 0\,\forall x$ it is also not a quasimeasure. Non-computability and non-dominance of W have already been shown in (ii). □

7. Conclusions

We investigated a natural strengthening of Solomonoff's famous convergence theorem, the latter stating that with probability 1 (w.p.1) the prediction of a universal semimeasure M converges to the true computable distribution μ ($M \overset{w.p.1}{\longrightarrow} \mu$). We answered partially negative the question of whether convergence also holds individually for all Martin-Löf (M.L.) random sequences ($\exists M : M \overset{M.L.}{\not\to} \mu$). We constructed random sequences α for which there exist universal semimeasures on which convergence fails. Multiplicative dominance of M is the key property to show convergence w.p.1. Dominance over all measures is also satisfied by the restricted mixture W over all quasimeasures. We showed that W converges to μ on all M.L.-random sequences by exploiting the incomputable mixture D over all measures. For $D \overset{M.L.}{\longrightarrow} \mu$ we achieved a (weak) convergence rate; for $W \overset{M.L.}{\longrightarrow} D$ and $W/D \overset{M.L.}{\longrightarrow} 1$ only an asymptotic result. The convergence rate properties w.p.1. of D and W are as excellent as for M.

We do not know whether $D/\mu \overset{M.L.}{\longrightarrow} 1$ holds. We also do not know the convergence rate for $W \overset{M.L.}{\longrightarrow} D$, and the current bound for $D \overset{M.L.}{\longrightarrow} \mu$ is double exponentially worse than for $M \overset{w.p.1}{\longrightarrow} \mu$. A minor question is whether D is approximable (which is unlikely). Finally there could still exist *universal* semimeasures M (dominating all enumerable semimeasures) for which M.L.-convergence holds ($\exists M : M \overset{M.L.}{\longrightarrow} \mu$?). In the case where they exist, we expect them to have particularly interesting additional structure and properties. While most results in algorithmic information theory are independent of the choice of the underlying universal Turing machine (UTM) or universal semimeasure (USM), there are also results which depend on this choice. For instance, one can show that $\{(x, n) : K_U(x) \leq n\}$ is tt-complete for some U, but not tt-complete for others [10]. A potential U dependence also occurs for predictions based on monotone complexity [5]. It could lead to interesting insights to identify a class of "natural" UTMs/USMs which have a variety of favorable properties. A more moderate approach may be to consider classes $\mathcal{C}_i$ of UTMs/USMs satisfying certain properties $\mathcal{P}_i$ and showing that the intersection $\cap_i \mathcal{C}_i$ is not empty.

Another interesting and potentially fruitful approach to the convergence problem at hand is to consider other classes of semimeasures $\mathcal{M}$, define mixtures M over $\mathcal{M}$, and (possibly) generalized randomness concepts by using this M in Definition 5. Using this approach, in [3] it has been shown that convergence holds for a subclass of Bernoulli distributions if the class is dense, but fails if the class is gappy, showing that a denseness characterization of $\mathcal{M}$ could be promising in general.

Acknowledgements

We want to thank Alexey Chernov for his invaluable help. This work was partially supported by the Swiss National Science Foundation (SNF grants 2100-67712 and 200020-107616) and the Russian Foundation for Basic Research (RFBR grants N04-01-00427 and N02-01-22001). A shorter version appeared in the proceedings of the ALT 2004 conference [1].

References

[1] M. Hutter, An.A. Muchnik, Universal convergence of semimeasures on individual random sequences, in: Proc. 15th International Conf. on Algorithmic Learning Theory, ALT'04, Padova, in: LNAI, vol. 3244, Springer, Berlin, 2004, pp. 234–248.

[2] M. Hutter, Convergence and loss bounds for Bayesian sequence prediction, IEEE Transactions on Information Theory 49 (8) (2003) 2061–2067.

[3] M. Hutter, On the existence and convergence of computable universal priors, in: Proc. 14th International Conf. on Algorithmic Learning Theory, ALT'03, Sapporo, in: LNAI, vol. 2842, Springer, Berlin, 2003, pp. 298–312.

[4] M. Hutter, An open problem regarding the convergence of universal a priori probability, in: Proc. 16th Annual Conf. on Learning Theory, COLT'03, Washington, DC, in: LNAI, vol. 2777, Springer, Berlin, 2003, pp. 738–740.

[5] M. Hutter, Sequence prediction based on monotone complexity, in: Proc. 16th Annual Conf. on Learning Theory, COLT'03, Washington, DC, in: LNAI, vol. 2777, Springer, Berlin, 2003, pp. 506–521.

[6] M. Hutter, Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability, Springer, Berlin, 2005, 300 pages. http://www.idsia.ch/˜marcus/ai/uaibook.htm.

[7] L.A. Levin, On the notion of a random sequence, Soviet Mathematics Doklady 14 (5) (1973) 1413–1416.

[8] M. Li, P.M.B. Vitányi, An Introduction to Kolmogorov Complexity and its Applications, 2nd edition, Springer, Berlin, 1997.

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

[10] An.A. Muchnik, S.Y. Positselsky, Kolmogorov entropy in the context of computability theory, Theoretical Computer Science 271 (1–2) (2002) 15–35.

[11] J. Schmidhuber, Algorithmic theories of everything, Report IDSIA-20-00, IDSIA, Manno, Lugano, Switzerland, quant-ph/0011122, 2000.

[12] J. Schmidhuber, Hierarchies of generalized Kolmogorov complexities and nonenumerable universal measures computable in the limit, International Journal of Foundations of Computer Science 13 (4) (2002) 587–612.

[13] R.J. Solomonoff, A formal theory of inductive inference: Parts 1 and 2, Information and Control 7 (1964) 1–22; 224–254.

[14] R.J. Solomonoff, Complexity-based induction systems: Comparisons and convergence theorems, IEEE Transactions on Information Theory IT-24 (1978) 422–432.

[15] P.M.B. Vitányi, M. Li, Minimum description length induction, Bayesianism, and Kolmogorov complexity, IEEE Transactions on Information Theory 46 (2) (2000) 446–464.

[16] V.G. Vovk, On a randomness criterion, Soviet Mathematics Doklady 35 (3) (1987) 656–660.

[17] A.K. Zvonkin, L.A. Levin, The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms, Russian Mathematical Surveys 25 (6) (1970) 83–124.