Universal Convergence of Semimeasures on Individual Random Sequences*

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch http://www.idsia.ch/~marcus

Andrej Muchnik

Institute of New Technologies, 10 Nizhnyaya Radischewskaya
Moscow 109004, Russia muchnik@lpcs.math.msu.ru

July 23, 2004

Abstract

Solomonoff's central result on induction is that the posterior of a universal semimeasure M converges rapidly and with probability 1 to the true sequence generating posterior μ, if the latter is computable. Hence, M is eligible as a universal sequence predictor in 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 for all 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; posterior convergence; Martin-Löf randomness; quasimeasures.

*This work was partially supported by the Swiss National Science Foundation (SNF grant 2100-67712.02) and the Russian Foundation for Basic Research (RFBR grants N04-01-00427 and N02-01-22001).

1 Introduction

A sequence prediction task is defined as to predict the next symbol x_n from an observed sequence $x{=}x_1...x_{n-1}$. The key concept to attack general prediction problems is Occam's razor, and to a less extent Epicurus' principle of multiple explanations. The former/latter may be interpreted as to keep the simplest/all theories consistent with the observations $x_1...x_{n-1}$ and to use these theories to predict x_n. Solomonoff [Sol64, Sol78] formalized and combined both principles in his universal prior M which assigns high/low probability to simple/complex environments x, hence implementing Occam and Epicurus. Formally it is a mixture of all enumerable semimeasures. An abstract characterization of M by Levin [ZL70] is that M is a universal enumerable semimeasure in the sense that it multiplicatively dominates all enumerable semimeasures.

Solomonoff's [Sol78] central result is that if the probability $\mu(x_n|x_1...x_{n-1})$ of observing x_n at time n, given past observations $x_1...x_{n-1}$ is a computable function, then the universal posterior $M_n := M(x_n|x_1...x_{n-1})$ converges (rapidly!) *with μ-probability 1* (w.p.1) for $n\to\infty$ to the true posterior $\mu_n{:=}\mu(x_n|x_1...x_{n-1})$, hence M represents a universal predictor in 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.... 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 [ML66, LV97]. 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 [Lev73]. Despite several attempts to solve this problem [Vov87, VL00, Hut03b], it remained open [Hut03c].

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 posterior M.L.-convergence of W to $\hat{D}$ and of $\hat{D}$ to μ. The Hellinger distance measuring closeness of two posterior 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 posterior 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 universal semimeasure M and an μ-M.L.-random sequence on which M does not converge to μ for some computable μ. 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 posterior of $\hat{D}$ M.L.-converges to μ. Finally, in Section 6 we show that W is non-universal and asymptotically M.L.-converges to $\hat{D}$. Section 7 contains discussion and outlook.

2 Notation & Universal Semimeasures M

Strings. Let $i,k,n,t\in I\!N=\{1,2,3,...\}$ 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 denote strings x of length $\ell(x)=n$ by $x=x_1x_2...x_n\in\mathcal{X}^n$ with $x_t\in\mathcal{X}$ and further abbreviate $x_{k:n}:=x_kx_{k+1}...x_{n-1}x_n$ for $k\leq n$, and $x_{<n}:=x_1...x_{n-1}$, 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, $\mathcal{X}^*\cong I\!N\cup\{0\}$).

Sets and functions. $\mathbb{Q}$, $I\!R$, $I\!R_+:=[0,\infty)$ 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 $f(n)\stackrel{n\to\infty}{\longrightarrow}g(n)$ and say f converges to g, without implying that $\lim_{n\to\infty}g(n)$ itself exists. We write $f(x)\stackrel{\times}{\leq}g(x)$ for $f(x)=O(g(x))$ and $f(x)\stackrel{+}{\leq}g(x)$ for $f(x)\leq g(x)+O(1)$.

Computability. A function $f:\mathcal{S}\to I\!R\cup\{\infty\}$ is said to be enumerable (or lower semi-computable) 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 semi-computable) 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 $g:\mathcal{S}\times I\!N\to I\!R$ with $\lim_{n\to\infty}g(x,n)=f(x)$. The set of enumerable functions is recursively enumerable.

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 $y\in\mathcal{X}^*$ [LV97]. $K(x):=K(x|\epsilon)$. 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 $(f_i)_{i=1}^n$ is an enumeration of all enumerable functions, we define $K(f_i)=K(i)$. We only need the following elementary properties: The co-enumerability of K, the upper bounds $K(x|\ell(x))\stackrel{+}{\leq}\ell(x)\log|\mathcal{X}|$ and $K(n)\stackrel{+}{\leq}2\log n$, and $K(x|y)\stackrel{+}{\leq}K(x)$, subadditivity $K(x)\stackrel{+}{\leq}K(x,y)\stackrel{+}{\leq}K(y)+K(x|y)$, and information non-increase $K(f(x))\stackrel{+}{\leq}K(x)+K(f)$ for recursive $f:\mathcal{X}^*\to\mathcal{X}^*$.

We need the concepts of (universal) (semi)measures for strings [ZL70].

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 posterior 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*

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

From now on we consider the (in a sense) largest class $\mathcal{M}$ which is relevant from a constructive point of view (but see [Sch02, Hut03b] 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 [Sol64, Eq.(7)] defined the universal posterior $M(x|y)=M(xy)/M(y)$ 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 [ZL70] has shown that this M is a universal enumerable semimeasure. Another possible definition of M is as a (Bayes) mixture [Sol64, ZL70, Sol78, LV97, Hut03b]: $\tilde M(x)=\sum_{\nu\in\mathcal{M}}2^{-K(\nu)}\nu(x)$, where $K(\nu)$ is the length of the shortest program computing function ν. Levin [ZL70] 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

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

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 other countable classes 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 Posterior Convergence with Probability 1

The following convergence results for M are well-known [Sol78, LV97, Hut03a].

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

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

The first convergence in difference is Solomonoff's [Sol78] celebrated convergence result. The second convergence in ratio has first been derived by Gács [LV97]. 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 $\mu(0_{1:n})=\prod_{t=1}^n(1-\frac{1}{2}t^{-3})\stackrel{n\to\infty}{\longrightarrow} c=0.450...>0$, i.e. $0_{1:\infty}$ is μ-random. On the other hand, one can show that $M(0_{<n})=O(1)$ and $M(0_{<n}1)\stackrel{\times}{=}2^{-K(n)}$, which implies $\frac{M(1|0_{<n})}{\mu(1|0_{<n})}\stackrel{\times}{=}n^3\cdot 2^{-K(n)}\stackrel{\times}{\geq} n\to\infty$ for $n\to\infty$ ($K(n)\stackrel{+}{\leq}2\log n$).

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 $f:\mathcal{X}^n\to I\!R$, $\mathbf{E}_\rho[f]=\sum'_{x_{1:n}}\rho(x_{1:n})f(x_{1:n})$, where $\sum'$ sums over all $x_{1:n}$ for which $\rho(x_{1:n})\neq 0$. Using $\sum'$ instead $\sum$ is 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:*

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

where $\mathbf{E}$ means expectation w.r.t. μ.

The $\ln w^{-1}$-bounds on the first and second expression have first been derived in [Hut03a], 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 $x_1x_2...$ 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:

$$\mathbf{P}[\#\{t:h_t\geq\varepsilon\}\geq\tfrac{1}{\varepsilon}(\ln w^{-1}+c)] \;\leq\; \mathbf{P}[\textstyle\sum_t h_t\geq\ln w^{-1}+c]$$

$$=\; \mathbf{P}[\exp(\tfrac{1}{2}\textstyle\sum_t h_t)\geq \mathrm{e}^{c/2}w^{-1/2}] \;\leq\; \sqrt{w}\mathbf{E}[\exp(\tfrac{1}{2}\textstyle\sum_t h_t)]\mathrm{e}^{-c/2} \;\leq\; \mathrm{e}^{-c/2}.$$

Proof. We use the abbreviations $\rho_t=\rho(x_t|x_{<t})$ and $\rho_{1:n}=\rho_1\cdot...\cdot\rho_n=\rho(x_{1:n})$ for $\rho\in\{\mu,\nu,R,N,...\}$ 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 [Vov87, Thm.1]. For discrete (semi)measures p and q with $\sum_i p_i=1$ and $\sum_i q_i\leq 1$ it holds:

$$\sum_i\sqrt{p_iq_i} \;\leq\; 1-\tfrac{1}{2}\sum_i(\sqrt{p_i}-\sqrt{q_i})^2 \;\leq\; \exp[-\tfrac{1}{2}\sum_i(\sqrt{p_i}-\sqrt{q_i})^2]. \qquad (3)$$

The first inequality is obvious after multiplying out the second expression. The second inequality follows from $1-x\leq\mathrm{e}^{-x}$. Vovk [Vov87] 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

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

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 $\sum_i p_i^{1-\kappa}q_i^\kappa\leq\exp(-\frac{1}{2}\sum_i(p_i^\kappa-q_i^\kappa)^{1/\kappa})$. $\square$

One can show that the constant $\frac{1}{2}$ in Lemma 4 can essentially not been 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 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 [Lev73]:

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: [Vov87] contains a (non-improvable?) result which is slightly too weak to imply M.L.-convergence, [LV97, Thm.5.2.2] and [VL00, Thm.10] contain an erroneous proof for M.L.-convergence, and [Hut03b] proves a theorem indicating that the answer may be hard and subtle (see [Hut03b] 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*

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

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. The proof idea is to construct an enumerable (semi)measure ν such that ν dominates M on some μ-random sequence α, but $\nu(\alpha_n|\alpha_{<n})\not\to\mu(\alpha_n|\alpha_{<n})$. Then we mix M to ν to make ν universal, but with larger contribution from ν, in order to preserve non-convergence. There is also non-constructive proof showing that an arbitrary small contamination with ν can lead to non-convergence. We only present the constructive proof.

Proof. We consider binary alphabet $\mathcal{X}=\{0,1\}$ only. Let $\mu(x)=\lambda(x):=2^{-\ell(x)}$ be the uniform measure. We define the sequence α as the (in a sense) lexicographically first (or equivalently left-most in the tree of sequences) λ.M.L.-random sequence. Formally we define α, inductively in $n=1,2,3,...$ 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

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

Let M^t with $t=1,2,3,...$ be computable approximations of M, which enumerate M, i.e. $M^t(x)\nearrow M(x)$ for $t\to\infty$. W define α^t like α but with M replaced by M^t in the definition. $M^t\nearrow M$ implies $\alpha^t\nearrow\alpha$ (lexicographically increasing). We define an enumerable semimeasure ν as follows:

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

where $<$ is the lexicographical ordering on sequences. 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

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

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

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

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:

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

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

$$\begin{aligned} M'(\alpha_n|\alpha_{<n}) &= \frac{(1-\gamma)\nu(\alpha_{1:n})+\gamma M(\alpha_{1:n})}{(1-\gamma)\nu(\alpha_{<n})+\gamma M(\alpha_{<n})} \overset{\substack{M(\alpha_{<n})\le 2^{-n+1} \text{ and } M(\alpha_{1:n})\ge 0\\ \downarrow}}{\geq} \frac{(1-\gamma)\nu(\alpha_{1:n})}{(1-\gamma)\nu(\alpha_{<n})+\gamma 2^{-n+1}} \\ &\underset{\substack{\uparrow\\ \nu(\alpha_{<n})=\nu(\alpha_{1:n})}}{=} \frac{1-\gamma}{1-\gamma+\gamma 2^{-n+1}/\nu(\alpha_{1:n})} \underset{\substack{\uparrow\\ \nu(\alpha_{1:n})\ge 2^{-n-1}}}{\geq} \frac{1-\gamma}{1+3\gamma} > \frac{1}{2}. \end{aligned}$$

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. $\square$

A converse of Theorem 6 can also be shown:

Theorem 7 (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 and the next section we give a positive answer to the question of posterior M.L.-convergence to μ. We consider general finite alphabet $\mathcal{X}$.

Theorem 8 (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 posteriors converge to each other:*

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

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 posterior converges to μ, but this measure is not enumerable, only approximable. For proving Theorem 8 we first define an intermediate measure D as a mixture over all computable measures, which is not even approximable. Based on Lemmas 4,9,10, Proposition 11 shows that D M.L.-converges to μ. We then define the concept of quasimeasures and an enumerable semimeasure W as a mixture over all enumerable quasimeasures. Proposition 12 shows that W M.L.-converges to D. Theorem 8 immediately follows from Propositions 11 and 12.

Lemma 9 (Hellinger Chain) *Let $h(p,q):=\sum_{i=1}^N(\sqrt{p_i}-\sqrt{q_i})^2$ be the Hellinger distance between $p=(p_i)_{i=1}^N\in I\!R_+^N$ and $q=(q_i)_{i=1}^N\in I\!R_+^N$. Then*

$$\begin{array}{lllll} i) & \textit{for } p,q,r\in I\!\!R^N_+ & h(p,q) & \leq & (1+\beta)\,h(p,r)+(1+\beta^{-1})\,h(r,q), \;\; \textit{any } \beta>0 \\ ii) & \textit{for } p^1,...,p^m\in I\!\!R^N_+ & h(p^1,p^m) & \leq & 3\sum_{k=2}^m k^2\,h(p^{k-1},p^k) \end{array}$$

Proof. (i) For any $x,y\in I\!\!R$ and $\beta>0$ we have $(x+y)^2\leq(1+\beta)x^2+(1+\beta^{-1})y^2$. Inserting $x=\sqrt{p_i}-\sqrt{r_i}$ and $y=\sqrt{r_i}-\sqrt{q_i}$ and summing over i proves (i).

(ii) Apply (i) for the triples (p^k,p^{k+1},p^m) for and in order of $k=1,2,...,m-2$ with $\beta=\beta_k=k(k+1)$ and finally use $\prod_{j=1}^{k-2}(1+\beta_j^{-1})\leq \mathrm{e}\leq 3$. □

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-submartingale $\bar H$. The universal μ-submartingale M/μ multiplicatively dominates all enumerable submartingales (and hence $\bar H$). Since $M/\mu\leq 2^{d_\mu(\omega)}$, this implies the desired bound $H(\omega)\stackrel{\times}{\leq}2^{d_\mu(\omega)}$ for individual ω. We give a self-contained direct proof, explicating all important constants.

Lemma 10 (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:*

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

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

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

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

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

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

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

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

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

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

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

Let $\mathcal{M}=\{\nu_1,\nu_2,...\}$ 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 posterior convergence of D to μ on μ-random sequences.

Proposition 11 (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:*

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

Combining (i) and (ii), using Lemma 9(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 10. The problem is that $\hat D$ is not computable, which we circumvent by joining with Lemma 9, bounds on $\sum_t h_t(\hat\delta_{k-1},\hat\delta_k)$ for $k=k_0,k_0+1,...$.

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 $d_\mu(\omega)<$

∞ 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 10 then implies $\exp(\frac12 H(\omega))\stackrel{\times}{\leq}\varepsilon_{k_0}^{-1/2}\cdot 2^{K(\mu,H,\sqrt{\varepsilon}_{k_0})+d_\mu(\omega)}$. We bound

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

The first inequality holds, since k_0 is the index and hence a description of μ, and ε_* 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 $K(k_0)\stackrel{+}{\leq}2\log k_0$ and the fact that for each $i\leq k_0$ one bit suffices to specify (non)membership to J_{k_0}, i.e. $K(J_{k_0}|k_0)\stackrel{+}{\leq}k_0$. Putting everything together we get

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

(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 10 then implies $H^k(\omega)\stackrel{\times}{\leq}\frac{\varepsilon_k}{\varepsilon_O\varepsilon_{k_0}}\cdot 2^{K(\mu,H^k,\varepsilon_O\varepsilon_{k_0}/\varepsilon_k)+d_\mu(\omega)}$. Similarly as in (i) we can bound

$$K(\mu,H^k,\varepsilon_{k_0}/\varepsilon_O\varepsilon_k)\stackrel{+}{\leq}K(J_k|k)+K(k)+K(k_0)\stackrel{+}{\leq}k+2\log k+2\log k_0,\quad\text{hence}$$

$$H^k(\omega) \;\stackrel{\times}{\leq}\; \tfrac{\varepsilon_k}{\varepsilon_O\varepsilon_{k_0}}\cdot k_0^2k^2 2^k c_\omega \;\stackrel{\times}{=}\; k_0^8 2^{k_0}k^{-4}c_\omega,\quad\text{where}\quad c_\omega:=2^{d_\mu(\omega)}.$$

Chaining this bound via Lemma 9(ii) we get for $k_1>k_0$:

$$\begin{aligned}\sum_{t=1}^n h_t(\hat\delta_{k_0},\hat\delta_{k_1}) \;&\leq\; \sum_{t=1}^n 3\sum_{k=k_0+1}^{k_1}(k-k_0+1)^2 h_t(\hat\delta_{k-1},\hat\delta_k)\\ &\leq\; 3\sum_{k=k_0+1}^{k_1}k^2H^k(\omega) \;\stackrel{\times}{\leq}\; 3k_0^8 2^{k_0}c_\omega\sum_{k=k_0+1}^{k_1}k^{-2} \;\leq\; 3k_0^7 2^{k_0}c_\omega\end{aligned}$$

If we now take $k_1\to\infty$ we get $\sum_{t=1}^n h_t(\hat\delta_{k_0},\hat D)\stackrel{\times}{\leq}3k_0^7 2^{k_0+d_\mu(\omega)}$. Finally let $n\to\infty$. $\Box$

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 convert semimeasures ν to "quasimeasures" $\tilde{\nu}$ as follows:

$$\tilde{\nu}(x_{1:n}) := \nu(x_{1:n}) \quad \text{if} \quad \sum_{y_{1:n}} \nu(y_{1:n}) \;>\; 1-\frac{1}{n} \quad \text{and} \quad \nu(x_{1:n}) := 0 \quad \text{else.}$$

If the condition is violated for some n it is also violated for all larger n, hence with ν also $\tilde{\nu}$ is a semimeasure. $\tilde{\nu}$ is enumerable if ν is enumerable. So if ν_1,ν_2,... is an enumeration of all enumerable semimeasures, then $\tilde{\nu}_1$,$\tilde{\nu}_2$,... is an enumeration of all enumerable quasimeasures. The for us important properties are that $\tilde{\nu}_i \leq \nu_i$ -and- if ν_i is a measure, then $\tilde{\nu}_i \equiv \nu_i$, else $\nu_i(x)=0$ for sufficiently long x. We define the enumerable semimeasure

$$W(x) := \sum_{i=1}^{\infty} \varepsilon_i \tilde{\nu}_i(x), \quad \text{andnote that} \quad D(x) = \sum_{i\in J} \varepsilon_i \tilde{\nu}_i(x) \;\text{ with }\; J := \{i : \tilde{\nu}_i \text{ is measure}\}$$

with $\varepsilon_i = i^{-6}2^{-i}$ as before.

Proposition 12 (Convergence of enumerable W to incomputable D) *For every computable measure μ and for ω being μ-random, the following holds for $t\to\infty$:*

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

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 $\varepsilon'_x:=\frac{2c'c_\omega}{\varepsilon_{k_0}}k_x^{-4}2^{-k_x}\to 0$ for $\ell(x)\to\infty$. Inserting this into (8) we get

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

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

(iii) Let $a\in\mathcal{X}$. From $W(xa)\geq D(xa)$ $(W\geq D)$ and $W(x)\leq(1+\varepsilon'_x)D(x)$ (i) we get

$$\begin{array}{rcl} W(a|x) & \geq & (1+\varepsilon'_x)^{-1}D(a|x) \;\geq\; (1-\varepsilon'_x)D(a|x) \qquad \forall a\in\mathcal{X}, \quad \text{and} \\ 1-W(a|x) & \geq & \displaystyle\sum_{b\neq a}W(b|x) \;\geq\; (1-\varepsilon'_x)\sum_{b\neq a}D(b|x) \;=\; (1-\varepsilon'_x)(1-D(a|x)), \end{array}$$

where we used in the second line that W is a semimeasure and D proportional to a measure. Together this implies $|W(a|x)-D(a|x)|\leq\varepsilon'_x$. Since $\varepsilon'_x\to 0$ for μ-random x, this shows (iii). $h_x(W,D)\leq\varepsilon'_x$ can also be shown. □

Speed of convergence. The main convergence Theorem 8 now immediately follows from Propositions 11 and 12. We briefly remark on the convergence rate. Lemma 4 shows that $\mathbf{E}[\sum_t h_t(X,\mu)]$ is logarithmic in the index k_0 of μ for $X=M$ $(\ln w_{k_0}^{-1}\stackrel{\times}{=}\ln k_0)$, but linear for $X=[W,D,\delta_{k_0}]$ $(\ln\varepsilon_{k_0}\stackrel{\times}{=}k_0)$. 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 11 are linear and exponential in k_0, respectively. For $W\stackrel{M.L.}{\longrightarrow}D$ 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 13 (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.

7 Conclusions

We investigated a natural strengthening of Solomonoff's famous convergence theorem, the latter stating that with probability 1 (w.p.1) the posterior of a universal semimeasure M converges to the true computable distribution μ ($M \stackrel{w.p.1}{\longrightarrow} \mu$). We answered partially negative the question of whether convergence also holds individually for all Martin-Löf (M.L.) random sequences ($\exists M : M \stackrel{M.L.}{\not\longrightarrow} \mu$). 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 $D \stackrel{M.L.}{\longrightarrow} \mu$ we achieved a (weak) convergence rate; for $W \stackrel{M.L.}{\longrightarrow} D$ and $W/D \stackrel{M.L.}{\longrightarrow} 1$ 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 $D/\mu \stackrel{M.L.}{\longrightarrow} 1$ holds. We also don't know the convergence rate for $W \stackrel{M.L.}{\longrightarrow} D$, and the current bound for $D \stackrel{M.L.}{\longrightarrow} \mu$ is double exponentially worse than for $M \stackrel{w.p.1}{\longrightarrow} \mu$. 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 ($\exists M : M \stackrel{M.L.}{\longrightarrow} \mu\,?$). In case 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 [MP02]. A potential U dependence also occurs for predictions based on monotone complexity [Hut03d]. 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 [Hut03b] 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.

References

[Hut03a] M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003.

[Hut03b] M. Hutter. On the existence and convergence of computable universal priors. In *Proc. 14th International Conf. on Algorithmic Learning Theory (ALT-2003)*, volume 2842 of *LNAI*, pages 298–312, Berlin, 2003. Springer.

[Hut03c] M. Hutter. An open problem regarding the convergence of universal a priori probability. In *Proc. 16th Annual Conf. on Learning Theory (COLT-2003)*, volume 2777 of *LNAI*, pages 738–740, Berlin, 2003. Springer.

[Hut03d] M. Hutter. Sequence prediction based on monotone complexity. In *Proc. 16th Annual Conf. on Learning Theory (COLT-2003)*, Lecture Notes in Artificial Intelligence, pages 506–521, Berlin, 2003. Springer.

[Lev73] L. A. Levin. On the notion of a random sequence. *Soviet Mathematics Doklady*, 14(5):1413–1416, 1973.

[LV97] M. Li and P. M. B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2nd edition, 1997.

[ML66] P. Martin-Löf. The definition of random sequences. *Information and Control*, 9(6):602–619, 1966.

[MP02] An. A. Muchnik and S. Y. Positselsky. Kolmogorov entropy in the context of computability theory. *Theoretical Computer Science*, 271(1–2):15–35, 2002.

[Sch02] J. Schmidhuber. Hierarchies of generalized Kolmogorov complexities and nonenumerable universal measures computable in the limit. *International Journal of Foundations of Computer Science*, 13(4):587–612, 2002.

[Sol64] R. J. Solomonoff. A formal theory of inductive inference: Part 1 and 2. *Information and Control*, 7:1–22 and 224–254, 1964.

[Sol78] R. J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transaction on Information Theory*, IT-24:422–432, 1978.

[VL00] P. M. B. Vitányi and M. Li. Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, 46(2):446–464, 2000.

[Vov87] V. G. Vovk. On a randomness criterion. *Soviet Mathematics Doklady*, 35(3):656–660, 1987.

[ZL70] A. K. Zvonkin and 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):83–124, 1970.