On Generalized Computable Universal Priors and their Convergence*

Marcus Hutter

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

11 March 2005

Abstract

Solomonoff unified Occam's razor and Epicurus' principle of multiple explanations to one elegant, formal, universal theory of inductive inference, which initiated the field of algorithmic information theory. His central result is that the posterior of the universal semimeasure M converges rapidly to the true sequence generating posterior μ, if the latter is computable. Hence, M is eligible as a universal predictor in case of unknown μ. The first part of the paper investigates the existence and convergence of computable universal (semi)measures for a hierarchy of computability classes: recursive, estimable, enumerable, and approximable. For instance, M is known to be enumerable, but not estimable, and to dominate all enumerable semimeasures. We present proofs for discrete and continuous semimeasures. The second part investigates more closely the types of convergence, possibly implied by universality: in difference and in ratio, with probability 1, in mean sum, and for Martin-Löf random sequences. We introduce a generalized concept of randomness for individual sequences and use it to exhibit difficulties regarding these issues. In particular, we show that convergence fails (holds) on generalized-random sequences in gappy (dense) Bernoulli classes.

Keywords

Sequence prediction; Algorithmic Information Theory; Solomonoff's prior; universal probability; mixture distributions; posterior convergence; computability concepts; Martin-Löf randomness.

*A preliminary version appeared in the proceedings of the ALT 2003 conference [Hut03a]. This work was supported by SNF grant 2000-61847.00 to Jürgen Schmidhuber.

1 Introduction

All induction problems can be phrased as sequence prediction tasks. This is, for instance, obvious for time-series prediction, but also includes classification tasks. Having observed data x_t at times $t<n$, the task is to predict the t-th symbol x_t from sequence $x=x_1...x_{t-1}$. The key concept to attack general induction problems is *Occam's razor* (simplicity) principle, which says that "*Entities should not be multiplied beyond necessity.*" 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 $x_1...x_{t-1}$ and to use these theories to predict x_t. Kolmogorov (and others) defined the complexity of a string as the length of its shortest description on a universal Turing machine. The Kolmogorov complexity K is an excellent universal complexity measure, suitable for quantifying Occam's razor. There is (only) one disadvantage: K is not computable.

More precisely, a function f is said to be *recursive* (or *finitely computable*) if there exists a Turing machine that, given x, computes $f(x)$ and then halts. Some functions are not recursive but still *approximable* (or *limit-computable*) in the sense that there is a nonhalting Turing machine with an infinite (x-dependent) output sequence y_1,y_2,y_3,... and $\lim_{t\to\infty}y_t=f(x)$. If additionally the output sequence is monotone increasing/decreasing, then f is said to be *lower/upper semicomputable* (or *enumerable/co-enumerable*). Finally we call f *estimable* if some Turing machine, given x and a precision ε, finitely computes an ε-approximation of x. The major algorithmic property of K is that it is co-enumerable, but not recursive.

More suitable for predictions is Solomonoff's [Sol64, Sol78] *universal prior* $M(x)$ defined as the probability that the output of a universal monotone Turing machine U starts with string x when provided with fair coin flips on the input tape. $M(x)$ is enumerable and roughly $2^{-K(x)}$, hence implementing Occam's and also Epicurus' principles.

Assume now that strings x are sampled from a probability distribution μ, i.e. the probability of a string starting with x shall be $\mu(x)$. The probability of observing x_t at time t, given past observations $x_1...x_{t-1}$ is $\mu(x_t|x_1...x_{t-1})=\mu(x_1...x_t)/\mu(x_1...x_{t-1})$. Solomonoff's [Sol78] central result is that the universal posterior $M(x_t|x_1...x_{t-1})=M(x_1...x_t)/M(x_1...x_{t-1})$ converges rapidly to the true (objective) posterior probability $\mu(x_t|x_1...x_{t-1})$, if μ is an estimable measure, hence M can be used for predictions in case of unknown μ. One representation of M is as a $2^{-K(\mu)}$-weighted sum of *all* enumerable "defective" probability measures, called semimeasures. The (from this representation obvious) dominance $M(x)\geq 2^{-K(\mu)}\mu(x)$ for all enumerable μ is the central ingredient in the convergence proof.

Dominance and convergence immediately generalize to arbitrary weighted sums of (semi)measures of some arbitrary countable set $\mathcal{M}$. So what is so special about the class of all enumerable semimeasures $\mathcal{M}^{semi}_{enum}$? The larger we choose $\mathcal{M}$ the less restrictive is the essential assumption that $\mathcal{M}$ should contain the true distribution μ. Why not restrict to the still rather general class of estimable or recursive

(semi)measures? For *every* countable class $\mathcal{M}$ and $\xi_{\mathcal{M}}(x) := \sum_{\nu\in\mathcal{M}} w_\nu \nu(x)$ with $w_\nu > 0$, the important dominance $\xi_{\mathcal{M}}(x) \geq w_\nu \nu(x)\, \forall \nu \in \mathcal{M}$ is satisfied. The question is what properties $\xi_{\mathcal{M}}$ possesses. The distinguishing property of $\mathcal{M}^{semi}_{enum}$ is that $M = \xi_{\mathcal{M}^{semi}_{enum}}$ is itself an element of $\mathcal{M}^{semi}_{enum}$. On the other hand, for prediction, $\xi_{\mathcal{M}} \in \mathcal{M}$ is not by itself an important property. What matters is whether $\xi_{\mathcal{M}}$ is computable (in one of the senses we defined above) to avoid getting into the (un)realm of non-constructive math.

Our first contribution is to classify the existence of generalized computable (semi)measures. From [ZL70] we know that there is an enumerable semimeasure (namely M) that dominates all enumerable semimeasures in $\mathcal{M}^{semi}_{enum}$. We show that there is *no* estimable semimeasure that dominates all recursive measures (also mentioned in [ZL70]), and there is *no* approximable semimeasure that dominates all approximable measures. From this it follows that for a universal (semi)measure that at least satisfies the weakest form of computability, namely being approximable, the largest dominated class among the classes considered in this work is the class of enumerable semimeasures. This is the distinguishing property of $\mathcal{M}^{semi}_{enum}$ and M. This investigation was motivated by recent generalizations of Kolmogorov complexity and Solomonoff's prior by Schmidhuber [Sch00, Sch02].

The second contribution is to investigate more closely the types of convergence, possibly implied by universality: in difference and in ratio, with probability 1, in mean sum, and for Martin-Löf random sequences. We introduce a generalized concept of randomness for individual sequences and use it to exhibit difficulties regarding these issues. More concretely, we consider countable classes $\mathcal{M}$ of Bernoulli environments and show that $\xi_{\mathcal{M}}$ converges to μ on all generalized random sequences if and only if the class is dense.

Contents. In Section 2 we review various computability concepts and discuss their relation. In Section 3 we define the prefix Kolmogorov complexity K, the concept of (semi)measures, Solomonoff's universal prior M, and explain its universality. Section 4 summarizes Solomonoff's major convergence result, discusses general mixture distributions and the important universality property – multiplicative dominance. In Section 5 we define seven classes of (semi)measures based on four computability concepts. Each class may or may not contain a (semi)measures that dominates all elements of another class. We reduce the analysis of these 49 cases to four basic cases. Domination (essentially by M) is known to be true for two cases. The other two cases do not allow for domination. In Section 7 we investigate more closely the type of convergence implied by universality. We summarize the result on posterior convergence in difference ($\xi - \mu \to 0$) and improve the previous result [LV97] on the convergence in ratio $\xi/\mu \to 1$ by showing rapid convergence without use of martingales. In Section 8 we investigate whether convergence for all Martin-Löf random sequences could hold. We define a generalized concept of randomness for individual sequences and use it to show that proofs based on universality cannot decide this question. Section 9 concludes the paper.

Notation. We denote strings of length n over finite alphabet $\mathcal{X}$ by $x=x_1x_2...x_n$ with $x_t\in\mathcal{X}$ and further abbreviate $x_{1:n}:=x_1x_2...x_{n-1}x_n$ and $x_{<n}:=x_1...x_{n-1}$, ϵ for the empty string, $\ell(x)$ for the length of string x, and $\omega=x_{1:\infty}$ for infinite sequences. We write xy for the concatenation of string x with y. 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 $f(x)\trianglerighteq g(x)$ for $g(x)=O(f(x))$, i.e. if $\exists c>0:f(x)\geq cg(x)\forall x$.

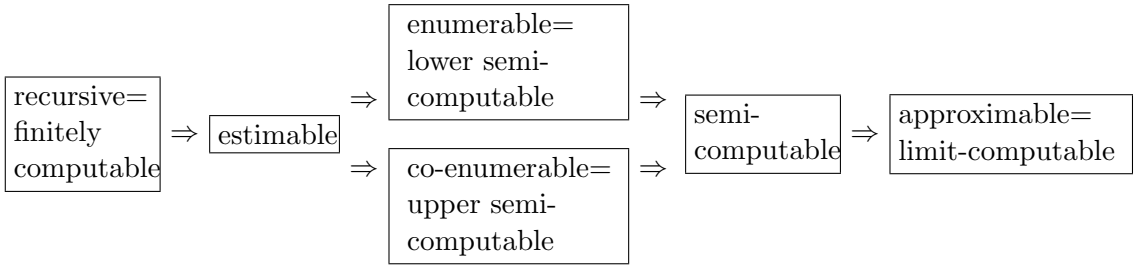
2 Computability Concepts

We define several computability concepts weaker than can be captured by halting Turing machines.

Definition 1 (Computable functions) *We consider functions $f:\mathbb{N}\to\mathbb{R}$:*

- *f is* recursive *or* finitely computable *iff there are Turing machines $T_{1/2}$ with output interpreted as natural numbers and $f(x)=\frac{T_1(x)}{T_2(x)}$,*
- *f is* approximable *or* limit-computable *iff $\exists$ recursive $\phi(\cdot,\cdot)$ with $\lim_{t\to\infty}\phi(x,t)=f(x)$.*
- *f is* enumerable *or* lower semicomputable *iff additionally $\phi(x,t)\leq\phi(x,t+1)$.*
- *f is* co-enumerable *or* upper semicomputable *iff $[-f]$ is lower semicomputable.*
- *f is* semicomputable *iff f is lower- or upper semicomputable.*
- *f is* estimable *iff f is lower- and upper semicomputable.*

If f is estimable we can finitely compute an ε-approximation of f by upper and lower semicomputing f and terminating when differing by less than ε. This means that there is a Turing machine which, given x and ε, finitely computes $\hat{y}\in\mathbb{Q}$ such that $|\hat{y}-f(x)|<\varepsilon$. Moreover it gives an interval estimate $f(x)\in[\hat{y}-\varepsilon,\hat{y}+\varepsilon]$. An estimable integer-valued function is recursive (take any $\varepsilon<\frac{1}{2}$). Note that if f is only approximable or semicomputable we can still come arbitrarily close to $f(x)$ but we cannot devise a terminating algorithm that produces an ε-approximation. In the case of lower/upper semicomputability we can at least finitely compute lower/upper bounds to $f(x)$. In case of approximability, the weakest computability form, even this capability is lost.



What we call *estimable/recursive/finitely computable* is often just called *computable*, but it makes sense to separate the concepts in this work, since finite computability is conceptually easier and some previous results have only been proved for this case. Sometimes we us the word *computable* generically for some of the computability forms of Definition 1.

3 The Universal Prior M

The prefix Kolmogorov complexity $K(x)$ is defined as the length of the shortest binary (prefix) program $p\in\{0,1\}^*$ for which a universal prefix Turing machine U (with binary program tape and $\mathcal{X}$ary output tape) outputs string $x\in\mathcal{X}^*$, and similarly $K(x|y)$ in case of side information y [Kol65, Lev74, Gác74, Cha75]:

$$K(x)=\min\{\ell(p):U(p)=x\},\qquad K(x|y)=\min\{\ell(p):U(p,y)=x\}$$

Solomonoff [Sol64, Eq.(7)] defined (earlier) the closely related quantity, the universal posterior $M(y|x)=M(xy)/M(x)$. The universal prior $M(x)$ can be defined as the probability that the output of a universal monotone Turing machine U starts with x when provided with fair coin flips on the input tape. Formally, M can be defined as

$$M(x)\;:=\sum_{p\,:\,U(p)=x*}2^{-\ell(p)} \tag{1}$$

where the sum is over minimal programs p for which U outputs a string starting with x. The so-called minimal programs are defined similarly to the prefix programs, but U need not to halt, which is indicated by the $*$. Minimal programs are those which are left to the input head in the moment when U wrote the last bit of x [LV97, Hut04]. Before we can discuss the stochastic properties of M we need the concept of (semi)measures for strings.

Definition 2 (Continuous (Semi)measures) *$\mu(x)$ denotes the probability that a sequence starts with string x. We call $\mu\geq 0$ a (continuous) semimeasure if $\mu(\epsilon)\leq 1$ and $\mu(x)\geq\sum_{a\in\mathcal{X}}\mu(xa)$, and a (probability) measure if equalities hold.*

The reason for calling μ with the above property a probability measure is that it satisfies Kolmogorov's axioms of probability in the following sense: The sample space is $\mathcal{X}^\infty$ with elements $\omega=\omega_1\omega_2\omega_3...\in\mathcal{X}^\infty$ being infinite sequences over alphabet $\mathcal{X}$. The set of events (the σ-algebra) is defined as the set generated from the cylinder sets $\Gamma_{x_{1:n}}:=\{\omega:\omega_{1:n}=x_{1:n}\}$ by countable union and complement. A probability measure μ is uniquely defined by giving its values $\mu(\Gamma_{x_{1:n}})$ on the cylinder sets, which we abbreviate by $\mu(x_{1:n})$. We will also call μ a measure, or even more loose a probability distribution.
We have $\sum_{a\in\mathcal{X}}M(xa)<M(x)$ because there are programs p that output x, not followed by any $a\in\mathcal{X}$. They just stop after printing x or continue forever without

any further output. Together with $M(\epsilon)=1$ this shows that M is a semimeasure, but *not* a probability measure. We can now state the fundamental property of M [ZL70, Sol78]:

Theorem 3 (Universality of M) *The universal prior M is an enumerable semimeasure that multiplicatively dominates all enumerable semimeasures in the sense that $M(x) \stackrel{\geq}{} 2^{-K(\rho)} \cdot \rho(x)$ for all enumerable semimeasures ρ. M is enumerable, but not estimable (nor recursive).*

The Kolmogorov complexity of a function like ρ is defined as the length of the shortest self-delimiting code of a Turing machine computing this function in the sense of Definition 1. Up to a multiplicative constant, M assigns higher probability to all x than any other computable probability distribution.

It is possible to normalize M to a true probability measure M_{norm} [Sol78, LV97] with dominance still being true, but at the expense of giving up enumerability (M_{norm} is still approximable). M is more convenient when studying algorithmic questions, but a true probability measure like M_{norm} is more convenient when studying stochastic questions.

4 Universal Sequence Prediction

In which sense does M incorporate Occam's razor and Epicurus' principle of multiple explanations? Since the shortest programs p dominate the sum in M, $M(x)$ is roughly equal to $2^{-K(x)}$ ($M(x)=2^{-K(x)+O(K(\ell(x))}$), i.e. M assigns high probability to simple strings. More useful is to think of x as being the observed history. We see from (1) that every program p consistent with history x is allowed to contribute to M (Epicurus). On the other hand, shorter programs give significantly larger contribution (Occam). How does all this affect prediction? If $M(x)$ describes our (subjective) prior belief in x, then $M(y|x):=M(xy)/M(x)$ must be our posterior belief in y. From the symmetry of algorithmic information $K(xy)\approx K(y|x)+K(x)$, and $M(x)\approx 2^{-K(x)}$ and $M(xy)\approx 2^{-K(xy)}$ we get $M(y|x)\approx 2^{-K(y|x)}$. This tells us that M predicts y with high probability iff y has an easy explanation, given x (Occam & Epicurus).

The above qualitative discussion should not create the impression that $M(x)$ and $2^{-K(x)}$ always lead to predictors of comparable quality. Indeed, in the online/incremental setting, $K(y)=O(1)$ invalidates the consideration above. The proof of (3) below, for instance, depends on M being a semimeasure and the chain rule being exactly true, neither of them is satisfied by $2^{-K(x)}$. See [Hut03b] for a detailed analysis.

Sequence prediction algorithms try to predict the continuation $x_t \in \mathcal{X}$ of a given sequence $x_1...x_{t-1}$. The following bound shows that M predicts computable sequences well:

$$\sum_{t=1}^{\infty}(1-M(x_t|x_{<t}))^2 \;\leq\; -\tfrac{1}{2}\sum_{t=1}^{\infty}\ln M(x_t|x_{<t}) \;=\; -\tfrac{1}{2}\ln M(x_{1:\infty}) \;\leq\; \tfrac{1}{2}\ln 2\cdot Km(x_{1:\infty}), \tag{2}$$

where the monotone complexity $Km(x_{1:\infty})=\min\{\ell(p):U(p)=x_{1:\infty}\}$ is defined as the length of the shortest (nonhalting) program computing $x_{1:\infty}$ [ZL70, Lev73]. In the first inequality we have used $(1-a)^2\leq -\frac{1}{2}\ln a$ for $0\leq a\leq 1$. In the equality we exchanged the sum with the logarithm and eliminated the resulting product by the chain rule. In the last inequality we used $M(x)\geq 2^{-Km(x)}$, which follows from (1) by dropping all terms in $\sum_p$ except for the shortest p computing x. If $x_{1:\infty}$ is a computable sequence, then $Km(x_{1:\infty})$ is finite, which implies $M(x_t|x_{<t})\to 1$ ($\sum_{t=1}^{\infty}(1-a_t)^2<\infty\Rightarrow a_t\to 1$). This means, that if the environment is a computable sequence (whichsoever, e.g. the digits of π or e in $\mathcal{X}$ary representation), after having seen the first few digits, M correctly predicts the next digit with high probability, i.e. it recognizes the structure of the sequence.

Assume now that the true sequence is drawn from a computable probability distribution μ, i.e. the true (objective) probability of $x_{1:t}$ is $\mu(x_{1:t})$. The probability of x_t given $x_{<t}$ hence is $\mu(x_t|x_{<t})=\mu(x_{1:t})/\mu(x_{<t})$. Solomonoff's [Sol78] central result is that M converges to μ. More precisely, for binary alphabet, he showed that

$$\sum_{t=1}^{\infty}\sum_{x_{<t}\in\{0,1\}^{t-1}}\mu(x_{<t})\Big(M(0|x_{<t})-\mu(0|x_{<t})\Big)^2 \;\leq\; \tfrac{1}{2}\ln 2\cdot K(\mu)+O(1) \;<\; \infty. \tag{3}$$

The infinite sum can only be finite if the difference $M(0|x_{<t})-\mu(0|x_{<t})$ tends to zero for $t\to\infty$ with μ-probability 1 (see Definition 10(i) and [Hut01] or Section 7 for general alphabet). This holds for *any* computable probability distribution μ. The reason for the astonishing property of a single (universal) function to converge to *any* computable probability distribution lies in the fact that the set of μ-random sequences differ for different μ. Past data $x_{<t}$ are exploited to get a (with $t\to\infty$) improving estimate $M(x_t|x_{<t})$ of $\mu(x_t|x_{<t})$.

The universality property (Theorem 3) is the central ingredient in the proof of (3). The proof involves the construction of a semimeasure ξ whose dominance is obvious. The hard part is to show its enumerability and equivalence to M. Let $\mathcal{M}$ be the (countable) set of all enumerable semimeasures and define

$$\xi(x) := \sum_{\nu\in\mathcal{M}} 2^{-K(\nu)}\nu(x). \tag{4}$$

Then dominance

$$\xi(x)\geq 2^{-K(\nu)}\nu(x) \quad \forall\,\nu\in\mathcal{M} \tag{5}$$

is obvious. Is ξ lower semicomputable? To answer this question one has to be more precise. Levin [ZL70] has shown that the set of *all* lower semicomputable semimeasures is enumerable (with repetitions). For this (ordered multi) set $\mathcal{M}=\mathcal{M}^{semi}_{enum}:=\{\nu_1,\nu_2,\nu_3,...\}$ and $K(\nu_i):=K(i)$ one can easily see that ξ is lower semicomputable. Finally proving $M(x)\trianglerighteq\xi(x)$ also establishes universality of M (see [Sol78, LV97] for details).

The advantage of ξ over M is that it immediately generalizes to arbitrary weighted sums of (semi)measures for arbitrary countable $\mathcal{M}$.

5 Universal (Semi)Measures

What is so special about the set of all enumerable semimeasures $\mathcal{M}^{semi}_{enum}$? The larger we choose $\mathcal{M}$ the less restrictive is the assumption that $\mathcal{M}$ should contain the true distribution μ, which will be essential throughout the paper. Why do not restrict to the still rather general class of estimable or recursive (semi)measures? It is clear that for every countable (multi)set $\mathcal{M}$, the universal or mixture distribution

$$\xi(x) := \xi_{\mathcal{M}}(x) := \sum_{\nu\in\mathcal{M}} w_\nu \nu(x) \quad \text{with} \quad \sum_{\nu\in\mathcal{M}} w_\nu \leq 1 \quad \text{and} \quad w_\nu > 0 \qquad (6)$$

dominates all $\nu\in\mathcal{M}$. This dominance is necessary for the desired convergence $\xi\to\mu$ similarly to (3). The question is what properties ξ possesses. The distinguishing property of $\mathcal{M}^{semi}_{enum}$ is that ξ is itself an element of $\mathcal{M}^{semi}_{enum}$. When concerned with predictions, $\xi_{\mathcal{M}}\in\mathcal{M}$ is not by itself an important property, but whether ξ is computable in one of the senses of Definition 1. We define

$$\begin{aligned} \mathcal{M}_1 \trianglerighteq \mathcal{M}_2 \;&:\Leftrightarrow\; \text{there is an element of } \mathcal{M}_1 \text{ that dominates all elements of } \mathcal{M}_2 \\ &:\Leftrightarrow\; \exists\rho\in\mathcal{M}_1\ \forall\nu\in\mathcal{M}_2\ \exists w_\nu>0\ \forall x:\ \rho(x)\geq w_\nu\nu(x). \end{aligned}$$

$\trianglerighteq$ is transitive (but not necessarily reflexive) in the sense that $\mathcal{M}_1\trianglerighteq\mathcal{M}_2\trianglerighteq\mathcal{M}_3$ implies $\mathcal{M}_1\trianglerighteq\mathcal{M}_3$ and $\mathcal{M}_0\supseteq\mathcal{M}_1\trianglerighteq\mathcal{M}_2\supseteq\mathcal{M}_3$ implies $\mathcal{M}_0\trianglerighteq\mathcal{M}_3$. For the computability concepts introduced in Section 2 we have the following proper set inclusions

$$\begin{array}{ccccccc} \mathcal{M}^{msr}_{rec} & \subset & \mathcal{M}^{msr}_{est} & \equiv & \mathcal{M}^{msr}_{enum} & \subset & \mathcal{M}^{msr}_{appr} \\ \cap & & \cap & & \cap & & \cap \\ \mathcal{M}^{semi}_{rec} & \subset & \mathcal{M}^{semi}_{est} & \subset & \mathcal{M}^{semi}_{enum} & \subset & \mathcal{M}^{semi}_{appr} \end{array}$$

where $\mathcal{M}^{msr}_c$ stands for the set of all probability measures of appropriate computability type $c\in\{$rec=recursive, est=estimable, enum=enumerable, appr=approximable$\}$, and similarly for semimeasures $\mathcal{M}^{semi}_c$. From an enumeration of a measure ρ one can construct a co-enumeration by exploiting $\rho(x_{1:n})=1-\sum_{y_{1:n}\neq x_{1:n}}\rho(y_{1:n})$. This shows that every enumerable measure is also co-enumerable, hence estimable, which proves the identity $\equiv$ above.

With this notation, Theorem 3 implies $\mathcal{M}^{semi}_{enum}\trianglerighteq\mathcal{M}^{semi}_{enum}$. Transitivity allows to conclude, for instance, that $\mathcal{M}^{semi}_{appr}\trianglerighteq\mathcal{M}^{msr}_{rec}$, i.e. that there is an approximable semimeasure that dominates all recursive measures.

The standard "diagonalization" way of proving $\mathcal{M}_1\not\trianglerighteq\mathcal{M}_2$ is to take an arbitrary $\mu\in\mathcal{M}_1$ and "increase" it to ρ such that $\mu\not\trianglerighteq\rho$ and show that $\rho\in\mathcal{M}_2$. There are 7×7 combinations of (semi)measures $\mathcal{M}_1$ with $\mathcal{M}_2$ for which $\mathcal{M}_1\trianglerighteq\mathcal{M}_2$ could be true or false. There are four basic cases, explicated in the following theorem, from which the other 49 combinations displayed in Table 5 follow by transitivity.

Theorem 4 (Universal (semi)measures) *A semimeasure ρ is said to be universal for $\mathcal{M}$ if it multiplicatively dominates all elements of $\mathcal{M}$ in the sense $\forall\nu\exists w_\nu>0:\rho(x)\geq w_\nu\nu(x)\forall x$. The following holds true:*

o) $\exists\rho:\{\rho\}\trianglerighteq\mathcal{M}$*: For every countable set of (semi)measures $\mathcal{M}$, there is a (semi)measure that dominates all elements of $\mathcal{M}$.*

i) $\mathcal{M}^{semi}_{enum}\trianglerighteq\mathcal{M}^{semi}_{enum}$*: The class of enumerable semimeasures* contains *a universal element.*

ii) $\mathcal{M}^{msr}_{appr}\trianglerighteq\mathcal{M}^{semi}_{enum}$*: There* is *an approximable measure that dominates all enumerable semimeasures.*

iii) $\mathcal{M}^{semi}_{est}\not\trianglerighteq\mathcal{M}^{msr}_{rec}$*: There is* no *estimable semimeasure that dominates all recursive measures.*

iv) $\mathcal{M}^{semi}_{appr}\not\trianglerighteq\mathcal{M}^{msr}_{appr}$*: There is* no *approximable semimeasure that dominates all approximable measures.*

Table 5 (Existence of universal (semi)measures) *The entry in row r and column c indicates whether there is an r-able (semi)measure ρ dominating the set $\mathcal{M}$ that contains all c-able (semi)measures, where r,$c\in\{$recurs, estimat, enumer, approxim$\}$. Enumerable measures are estimable. This is the reason why the enum. row and column in case of measures are missing. The superscript indicates from which part of Theorem 4 the answer follows. For the bold face entries directly, for the others using transitivity of $\trianglerighteq$.*

$\nwarrow$	$\mathcal{M}$	*semimeasure*				*measure*		
ρ	$\searrow$	*rec.*	*est.*	*enum.*	*appr.*	*rec.*	*est.*	*appr.*
s	*rec.*	no^{iii}	no^{iii}	no^{iii}	no^{iv}	no^{iii}	no^{iii}	no^{iv}
e	*est.*	no^{iii}	no^{iii}	no^{iii}	no^{iv}	$\mathbf{no^{iii}}$	no^{iii}	no^{iv}
m	*enum.*	yes^{i}	yes^{i}	$\mathbf{yes^{i}}$	no^{iv}	yes^{i}	yes^{i}	no^{iv}
i	*appr.*	yes^{i}	yes^{i}	yes^{i}	no^{iv}	yes^{i}	yes^{i}	$\mathbf{no^{iv}}$
m	*rec.*	no^{iii}	no^{iii}	no^{iii}	no^{iv}	no^{iii}	no^{iii}	no^{iv}
s	*est.*	no^{iii}	no^{iii}	no^{iii}	no^{iv}	no^{iii}	no^{iii}	no^{iv}
r	*appr.*	yes^{ii}	yes^{ii}	$\mathbf{yes^{ii}}$	no^{iv}	yes^{ii}	yes^{ii}	no^{iv}

If we ask for a universal (semi)measure that at least satisfies the weakest form of computability, namely being approximable, we see that the largest dominated set among the 7 sets defined above is the set of enumerable semimeasures. This is the reason why $\mathcal{M}^{semi}_{enum}$ plays a special role. On the other hand, $\mathcal{M}^{semi}_{enum}$ is not the largest set dominated by an approximable semimeasure, and indeed no such largest set exists. One may, hence, ask for "natural" larger sets $\mathcal{M}$. One such set,

namely the set of cumulatively enumerable semimeasures $\mathcal{M}_{\rm CEM}$, has recently been discovered by Schmidhuber [Sch00, Sch02], for which even $\xi_{\rm CEM}\in\mathcal{M}_{\rm CEM}$ holds. Theorem 4 also holds for *discrete (semi)measures* P defined as follows:

Definition 6 (Discrete (semi)measures) *$P(x)$ denotes the probability of $x\in\mathbb{N}$. We call $P:\mathbb{N}\to[0,1]$ a discrete (semi)measure if $\sum_{x\in\mathbb{N}}P(x)\stackrel{(<)}{=}1$.*

Theorem 4 (i) is Levin's major result [LV97, Thm.4.3.1 & Thm.4.5.1], and (ii) is due to Solomonoff [Sol78]. The proof of $\mathcal{M}^{semi}_{rec}\not\supseteq\mathcal{M}^{semi}_{rec}$ in [LV97, p249] contains minor errors and is not extensible to (iii), and the proof in [LV97, p276] only applies to infinite alphabet and not to the binary/finite case considered here. $\mathcal{M}^{semi}_{est}\not\supseteq\mathcal{M}^{semi}_{est}$ is mentioned in [ZL70] without proof. A direct proof of (iv) can be found in [Hut04]. Here, we reduce (iv) to (iii) by exploiting the following elementary fact (well-known for integer-valued functions, see e.g. [Sim77, p634]):

Lemma 7 (Approximable = H-estimable) *A function is approximable iff it is estimable with the help of the halting oracle.*

Proof. With H-computable we mean, computable with the help of the halting oracle, or equivalently, computable under extra input of the halting sequence $h=h_{1:\infty}\in\{0,1\}^\infty$, where $h_n=1:\Leftrightarrow U(n)$ halts.

Assume f is approximable, i.e. $\forall\varepsilon\exists y,m:R(m,y,\varepsilon)$, where relation $R(m,y,\varepsilon):=[\forall n\geq m:|f_n(x)-y|<\varepsilon]$ and recursive $f_n\to f$. Fix $\varepsilon>0$. Search (dovetail) for $m\in\mathbb{N}$ and y ($\in\frac{1}{2}\varepsilon\mathbb{Z}$ is sufficient) such that $R(m,y,\varepsilon)$=true. R is co-enumerable, hence H-decidable, hence y can be H-computed, hence f is H-estimable, since $f(x)=y\pm O(\varepsilon)$.

Now assume that f is H-estimable, i.e. $\exists T\in$TM $\forall\varepsilon,x:|T(x,\varepsilon,h)-f(x)|<\varepsilon$. Since h is co-enumerable, T and hence f are approximable. More formally, let $h^t_n=1:\Leftrightarrow U(n)$ halts within t steps. Then $g(x,\varepsilon):=T(x,\varepsilon,h)=T(x,\varepsilon,\lim_{t\to\infty}h^t)=\lim_{t\to\infty}T(x,\varepsilon,h^t)$ is approximable, where the exchange of limits holds, since T only reads $n_{x\varepsilon}<\infty$ bits of h and $h_{1:n_{x\varepsilon}}=h^t_{1:n_{x\varepsilon}}$ for sufficiently large t. □

6 Proof of Theorem 4

We first prove the theorem for discrete (semi)measures P (Definition 6), since it contains the essential ideas in a cleaner form. We then present the proof for continuous (semi)measures μ (Definition 2). We present proofs for binary alphabet $\mathcal{X}=\{0,1\}$ only. The proofs naturally generalize from binary to arbitrary finite alphabet. $\operatorname{argmin}_x f(x)$ is the x that minimizes $f(x)$. Ties are broken in an arbitrary but computable way (e.g. by taking the smallest x).

Proof (discrete case).
(o) $Q(x):=\sum_{P\in\mathcal{M}}w_PP(x)$ with $w_P>0$ obviously dominates all $P\in\mathcal{M}$ (with constant

w_P). With $\sum_P w_P = 1$ and all P being discrete (semi)measures also Q is a discrete (semi)measure.

(i) See [LV97, Thm.4.3.1].

(ii) Let P be the universal element in $\mathcal{M}^{semi}_{enum}$ and $\alpha := \sum_x P(x)$. We normalize P by $Q(x) := \frac{1}{\alpha} P(x)$. Since $\alpha \leq 1$ we have $Q(x) \geq P(x)$. Hence $Q \geq P \trianglerighteq \mathcal{M}^{semi}_{enum}$. As a ratio between two enumerable functions, Q is still approximable, hence $\mathcal{M}^{msr}_{appr} \trianglerighteq \mathcal{M}^{semi}_{enum}$.

(iii) Let $P \in \mathcal{M}^{semi}_{rec}$. We partition $\mathbb{N}$ into chunks $I_n := \{2^{n-1},...,2^n-1\}$ $(n \geq 1)$ of increasing size. With $x_n := \text{argmin}_{x \in I_n} P(x)$ we define $Q(x_n) := \frac{1}{n(n+1)} \forall n$ and $Q(x) := 0$ for all other x. Exploiting that a minimum is smaller than an average and that μ is a semimeasure, we get

$$P(x_n) = \min_{x \in I_n} P(x) \leq \frac{1}{|I_n|} \sum_{x \in I_n} P(x) \leq \frac{1}{|I_n|} = \frac{1}{2^{n-1}} = \frac{n(n+1)}{2^{n-1}} Q(x_n)$$

Since $\frac{n(n+1)}{2^{n-1}} \to 0$ for $n \to \infty$, P cannot dominate Q $(P \not\trianglerighteq Q)$. With P also Q is recursive. Since P was an arbitrary recursive semimeasure and Q is a recursive measure $(\sum Q(x) = \sum[\frac{1}{n(n+1)}] = \sum[\frac{1}{n} - \frac{1}{n+1}] = 1)$ this implies $\mathcal{M}^{semi}_{rec} \not\trianglerighteq \mathcal{M}^{msr}_{rec}$.

Assume now that there is an estimable semimeasure $S \trianglerighteq \mathcal{M}^{msr}_{rec}$. We construct a recursive semimeasure $P \trianglerighteq S$ as follows. Choose an initial $\varepsilon > 0$ and finitely compute an ε-approximation $\hat{S}$ of $S(x)$. If $\hat{S} > 2\varepsilon$ define $P(x) := \frac{1}{2}\hat{S}$, else halve ε and repeat the process. Since $S(x) > 0$ (otherwise it could not dominate, e.g. $T(x) := \frac{1}{x(x+1)} \in \mathcal{M}^{msr}_{rec}$) the loop terminates after finite time. So P is recursive. Inserting $\hat{S} = 2P(x)$ and $\varepsilon < \frac{1}{2}\hat{S} = P(x)$ into $|S(x) - \hat{S}| < \varepsilon$ we get $|S(x) - 2P(x)| < P(x)$, which implies $S(x) \geq P(x)$ and $S(x) \leq 3P(x)$. The former implies $\sum_x P(x) \leq \sum_x S(x) \leq 1$, i.e. P is a semimeasure. The latter implies $P \geq \frac{1}{3} S \trianglerighteq \mathcal{M}^{msr}_{rec}$. Hence P is a recursive semimeasure dominating all recursive measures, which contradicts what we have proven in the first half of (iii). Hence the assumption on S was wrong which establishes $\mathcal{M}^{semi}_{est} \not\trianglerighteq \mathcal{M}^{msr}_{rec}$.

(iv) From (iii) we know that $\mathcal{M}^{semi}_{est} \not\trianglerighteq \mathcal{M}^{msr}_{est}$. The proof and hence result remains valid under the halting oracle, i.e. $\mathcal{M}^{semi}_{H\text{-}est} \not\trianglerighteq \mathcal{M}^{msr}_{H\text{-}est}$. By Lemma 7, the H-estimable functions/(semi)measures coincide with the approximable functions/(semi)measures, hence $\mathcal{M}^{semi}_{appr} \not\trianglerighteq \mathcal{M}^{msr}_{appr}$. □

Proof (continuous case).
The major difference to the discrete case is that one also has to take care that $\rho(x) \overset{(>)}{=} \rho(x0) + \rho(x1)$, $x \in \{0,1\}^*$, is respected. On the other hand, the chunking $I_n := \{0,1\}^n$ is more natural here.

(o) $\rho(x) := \sum_{\nu \in \mathcal{M}} w_\nu \nu(x)$ with $w_\nu > 0$ obviously dominates all $\nu \in \mathcal{M}$ (with domination constant w_ν). With $\sum_\nu w_\nu = 1$ and all ν being (semi)measures also ρ is a (semi)measure.

(i) See [LV97, Thm.4.5.1].

(ii) Let ξ be a universal element in $\mathcal{M}^{semi}_{enum}$. We define [Sol78]

$$\xi_{norm}(x_{1:n}) \;:=\; \prod_{t=1}^{n} \frac{\xi(x_{1:t})}{\xi(x_{<t}0)+\xi(x_{<t}1)}.$$

By induction one can show that ξ_{norm} is a measure and that $\xi_{norm}(x)\geq\xi(x)\forall x$, hence $\xi_{norm}\geq\xi\trianglerighteq\mathcal{M}^{semi}_{enum}$. As a ratio of enumerable functions, ξ_{norm} is still approximable, hence $\mathcal{M}^{msr}_{appr}\trianglerighteq\mathcal{M}^{semi}_{enum}$.

(iii) Analogous to the discrete case we could start by recursively defining $x^*_k := \text{argmin}_{x_k}\mu(x^*_{<k}x_k)$ for $\mu\in\mathcal{M}^{semi}_{rec}$. See [Hut03a] for a proof along this line. Simpler is to directly consider $\mu\in\mathcal{M}^{semi}_{est}$ and to compute $x^*_{1:\infty}$ recursively by computing some ε-approximation $e(x_k|x^*_{<t})$ of $\mu(x_k|x^*_{<t})$ and define $x^*_k=\text{argmax}_{x_k}e(x_k|x^*_{<t})$, which implies $\mu(x^*_k|x^*_{<t})\leq\frac{1}{2}+\varepsilon$. Finally we define measure ρ by $\rho(x^*_{1:k})=1\forall k$ and $\rho(x)=0$ for all x that are not prefixes of $x^*_{1:\infty}$. Hence $\mu(x^*_{1:n})\leq(\frac{1}{2}+\varepsilon)^n=(\frac{1}{2}+\varepsilon)^n\rho(x^*_{1:n})$, which demonstrates that μ does not dominate ρ for $\varepsilon<\frac{1}{2}$. Since $\mu\in\mathcal{M}^{semi}_{est}$ was arbitrary and ρ is a recursive measure, this implies $\mathcal{M}^{semi}_{est}\not\trianglerighteq\mathcal{M}^{msr}_{rec}$.

(iv) Identical to discrete case. □

7 Posterior Convergence

We investigated in detail the computational properties of various mixture distributions ξ. A mixture $\xi_{\mathcal{M}}$ multiplicatively dominates all distributions in $\mathcal{M}$. We mentioned that dominance implies posterior convergence. In this section we present in more detail what dominance implies and what not.

Convergence of $\xi(x_t|x_{<t})$ to $\mu(x_t|x_{<t})$ with μ-probability 1 tells us that $\xi(x_t|x_{<t})$ is close to $\mu(x_t|x_{<t})$ for sufficiently large t on 'most' sequences $x_{1:\infty}$. It says nothing about the speed of convergence, nor whether convergence is true for any *particular* sequence (of measure 0). Convergence *in mean sum* defined below is intended to capture the rate of convergence, Martin-Löf randomness is 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 prior. Levin gave a characterization equivalent to Martin-Löf's original definition [Lev73]:

Theorem 8 (Martin-Löf random sequences) *A sequence $x_{1:\infty}$ is μ-Martin-Löf random (μ.M.L.) iff there is a constant c such that $M(x_{1:n})\leq c\cdot\mu(x_{1:n})$ for all n.*

An equivalent formulation for estimable μ is:

$$x_{1:\infty}\text{ is }\mu\text{.M.L.-random} \quad\Leftrightarrow\quad Km(x_{1:n}) = -\log\mu(x_{1:n})+O(1)\;\forall n \tag{7}$$

Theorem 8 follows from (7) by exponentiation, "using $2^{-Km}\approx M$" and noting that $M\trianglerighteq\mu$ follows from universality of M. Consider the special case of μ being a fair

coin, i.e. $\mu(x_{1:n})=2^{-n}$, then $x_{1:\infty}$ is M.L. random *iff* $Km(x_{1:n})=n+O(1)$, i.e. if $x_{1:n}$ is incompressible. For general μ, $-\log\mu(x_{1:n})$ is the length of the Shannon-Fano code of $x_{1:n}$, hence $x_{1:\infty}$ is μ.M.L.-random *iff* the Shannon-Fano code is optimal.

One can show that a μ.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 following generalization is natural when considering general Bayes mixtures ξ as in this work:

Definition 9 (μ/ξ-random sequences) *A sequence $x_{1:\infty}$ is called μ/ξ-random ($\mu.\xi.r.$) iff there is a constant c such that $\xi(x_{1:n})\leq c\cdot\mu(x_{1:n})$ for all n.*

Typically, ξ is a mixture over some $\mathcal{M}$ as defined in (6), in which case the reverse inequality $\xi(x)\trianglerighteq\mu(x)$ is also true (for all x). For finite $\mathcal{M}$ or if $\xi\in\mathcal{M}$, the definition of μ/ξ-randomness depends only on $\mathcal{M}$, and not on the specific weights w_ν used in ξ. For $\mathcal{M}=\mathcal{M}^{semi}_{enum}$, μ/ξ-randomness is just μ.M.L.-randomness. The larger $\mathcal{M}$, the more patterns are recognized as nonrandom. Roughly speaking, those regularities characterized by some $\nu\in\mathcal{M}$ are recognized by μ/ξ-randomness, i.e. for $\mathcal{M}\subset\mathcal{M}^{semi}_{enum}$ some μ/ξ-random strings may not be M.L. random. Other randomness concepts, e.g. those by Schnorr, Ko, van Lambalgen, Lutz, Kurtz, von Mises, Wald, and Church (see [Wan96, Lam87, Sch71]), could possibly also be characterized in terms of μ/ξ-randomness for particular choices of $\mathcal{M}$.

A classical (nonrandom) real-valued sequence a_t is defined to converge to a_*, short $a_t\to a_*$ if $\forall\varepsilon\exists t_0\forall t\geq t_0:|a_t-a_*|<\varepsilon$. We are interested in convergence properties of random sequences $z_t(\omega)$ for $t\to\infty$ (e.g. $z_t(\omega)=\xi(\omega_t|\omega_{<t})-\mu(\omega_t|\omega_{<t})$). We denote μ-expectations by $\mathbf{E}$. The expected value of a function $f:\mathcal{X}^t\to\mathbb{R}$, dependent on $x_{1:t}$, independent of $x_{t+1:\infty}$, and possibly undefined on a set of μ-measure 0, is $\mathbf{E}[f]=\sum'_{x_{1:t}\in\mathcal{X}^t}\mu(x_{1:t})f(x_{1:t})$. The prime denotes that the sum is restricted to $x_{1:t}$ with $\mu(x_{1:t})\neq 0$. Similarly we use $\mathbf{P}[..]$ to denote the μ-probability of event $[..]$. We define four convergence concepts for random sequences.

Definition 10 (Convergence of random sequences) *Let $z_1(\omega),z_2(\omega),...$ be a sequence of real-valued random variables. z_t is said to converge for $t\to\infty$ to (random variable) z_**

i) with probability 1 (w.p.1) $:\Leftrightarrow \mathbf{P}[\{\omega:z_t\to z_*\}]=1$,

ii) in mean sum (i.m.s.) $:\Leftrightarrow \sum_{t=1}^{\infty}\mathbf{E}[(z_t-z_*)^2]<\infty$,

iii) for every μ-Martin-Löf random sequence (μ.M.L.) $:\Leftrightarrow$
$\forall\omega$: *If* $[\exists c\forall n:M(\omega_{1:n})\leq c\mu(\omega_{1:n})]$ *then* $z_t(\omega)\to z_*(\omega)$ *for* $t\to\infty$,

iv) for every μ/ξ-random sequence ($\mu.\xi.r.$) $:\Leftrightarrow$
$\forall\omega$: *If* $[\exists c\forall n:\xi(\omega_{1:n})\leq c\mu(\omega_{1:n})]$ *then* $z_t(\omega)\to z_*(\omega)$ *for* $t\to\infty$.

In statistics, (i) is the "default" characterization of convergence of random sequences. Convergence i.m.s. (ii) is very strong: it provides a rate of convergence in the sense that the expected number of times t in which z_t deviates more than ε from z_* is finite and bounded by c/ε^2 and the probability that the number of ε-deviations exceeds $\frac{c}{\varepsilon^2\delta}$ is smaller than δ, where $c:=\sum_{t=1}^{\infty}\mathbf{E}[(z_t-z_*)^2]$. Nothing can be said for *which* t these deviations occur. If, additionally, $|z_t-z_*|$ were monotone decreasing, then $|z_t-z_*|=o(t^{-1/2})$ could be concluded. (iii) uses Martin-Löf's notion of randomness of *individual* sequences to define convergence M.L. Since this work deals with general Bayes mixtures ξ, we generalized in (iv) the definition of convergence M.L. based on M to convergence $\mu.\xi$.r. based on ξ in a natural way. One can show that convergence i.m.s. implies convergence w.p.1. Also convergence M.L. implies convergence w.p.1. Universality of ξ implies the following posterior convergence results:

Theorem 11 (Convergence of ξ to μ) *Let there be sequences $x_1x_2...$ over a finite alphabet $\mathcal{X}$ drawn with probability $\mu(x_{1:n})\in\mathcal{M}$ for the first n symbols, where μ is a measure and $\mathcal{M}$ a countable set of (semi)measures. The universal/mixture posterior probability $\xi(x_t|x_{<t})$ of the next symbol x_t given $x_{<t}$ is related to the true posterior probability $\mu(x_t|x_{<t})$ in the following way:*

$$\sum_{t=1}^{n}\mathbf{E}\left[\left(\sqrt{\tfrac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})}}-1\right)^2\right] \leq \sum_{t=1}^{n}\mathbf{E}\left[\sum_{x'_t}\left(\sqrt{\xi(x'_t|x_{<t})}-\sqrt{\mu(x'_t|x_{<t})}\right)^2\right] \leq \ln w_\mu^{-1} < \infty$$

where w_μ is the weight (6) of μ in ξ.

Theorem 11 implies

$$\sqrt{\xi(x'_t|x_{<t})}\to\sqrt{\mu(x'_t|x_{<t})} \text{ for any } x'_t \text{ and } \sqrt{\tfrac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})}}\to 1, \text{ both i.m.s. for } t\to\infty.$$

The latter strengthens the result $\xi(x_t|x_{<t})/\mu(x_t|x_{<t})\to 1$ w.p.1 derived by Gács [LV97, Thm.5.2.2] in that it also provides the "speed" of convergence.

Note also the subtle difference between the two convergence results. For *any* sequence $x'_{1:\infty}$ (possibly constant and not necessarily μ-random), $\mu(x'_t|x_{<t})-\xi(x'_t|x_{<t})$ converges to zero w.p.1 (referring to $x_{1:\infty}$), but no statement is possible for $\xi(x'_t|x_{<t})/\mu(x'_t|x_{<t})$, since $\liminf\mu(x'_t|x_{<t})$ could be zero. On the other hand, if we stay *on*-sequence ($x'_{1:\infty}=x_{1:\infty}$), we have $\xi(x_t|x_{<t})/\mu(x_t|x_{<t})\to 1$ w.p.1 (whether $\inf\mu(x_t|x_{<t})$ tends to zero or not does not matter). Indeed, it is easy to give an example where $\xi(x'_t|x_{<t})/\mu(x'_t|x_{<t})$ diverges. If we choose

$$\mathcal{M}=\{\mu_1,\mu_2\},\quad \mu\equiv\mu_1,\quad \mu_1(1|x_{<t})=\tfrac{1}{2}t^{-3}\quad\text{and}\quad \mu_2(1|x_{<t})=\tfrac{1}{2}t^{-2}$$

the contribution of μ_2 to ξ causes ξ to fall off like $\mu_2\sim t^{-2}$, much slower than $\mu\sim t^{-3}$ causing the quotient to diverge:

$$\begin{aligned}
\mu_1(0_{1:n}) &= \prod_{t=1}^{n}(1-\tfrac{1}{2}t^{-3}) \stackrel{n\to\infty}{\longrightarrow} c_1 = 0.450... > 0 \;\Rightarrow\; 0_{1:\infty} \text{ is a } \mu\text{-random sequence,}\\
\mu_2(0_{1:n}) &= \prod_{t=1}^{n}(1-\tfrac{1}{2}t^{-2}) \stackrel{n\to\infty}{\longrightarrow} c_2 = 0.358... > 0 \;\Rightarrow\; \xi(0_{1:n}) \to w_1c_1 + w_2c_2 =: c_\xi > 0\\
\xi(0_{<t}1) &= w_1\mu_1(1|0_{<t})\mu_1(0_{<t}) + w_2\mu_2(1|0_{<t})\mu_2(0_{<t}) \to \tfrac{1}{2}w_2c_2t^{-2}
\end{aligned}$$

$$\Rightarrow \quad \xi(1|0_{<t}) = \frac{\xi(0_{<t}1)}{\xi(0_{<t})} \to \frac{w_2c_2}{2c_\xi}t^{-2} \quad \Rightarrow \quad \frac{\xi(1|0_{<t})}{\mu(1|0_{<t})} \to \frac{w_2c_2}{c_\xi}t \to \infty \quad \text{diverges.}$$

Proof. For a probability distribution $y_i \geq 0$ with $\sum_i y_i = 1$ and a semi-distribution $z_i \geq 0$ with $\sum_i z_i \leq 1$ and $i = \{1,...,N\}$, the Hellinger distance $h(\vec{y},\vec{z}) := \sum_i(\sqrt{y_i}-\sqrt{z_i})^2$ is upper bounded by the relative entropy $d(\vec{y},\vec{z}) = \sum_i y_i \ln \frac{y_i}{z_i}$ (and $0\ln\frac{0}{z} := 0$). This can be seen as follows: For arbitrary $0 \leq y \leq 1$ and $0 \leq z \leq 1$ we define

$$\begin{aligned}
f(y,z) &:= y\ln\frac{y}{z} - (\sqrt{y}-\sqrt{z})^2 + z - y = 2yg(\sqrt{z/y})\\
\text{with} \quad g(t) &:= -\ln t + t - 1 \geq 0.
\end{aligned}$$

This shows $f \geq 0$, and hence $\sum_i f(y_i,z_i) \geq 0$, which implies

$$\sum_i y_i \ln\frac{y_i}{z_i} - \sum_i(\sqrt{y_i}-\sqrt{z_i})^2 \geq \sum_i y_i - \sum_i z_i \geq 1 - 1 = 0.$$

The (conditional) μ-expectations of a function $f : \mathcal{X}^t \to \mathbb{R}$ are defined as

$$\mathbf{E}[f] = \sum_{x_{1:t}\in\mathcal{X}^t}{}' \mu(x_{1:t})f(x_{1:t}) \quad \text{and} \quad \mathbf{E}_t[f] := \mathbf{E}[f|x_{<t}] = \sum_{x_t\in\mathcal{X}}{}' \mu(x_t|x_{<t})f(x_{1:t}),$$

where $\sum'$ sums over all x_t or $x_{1:t}$ for which $\mu(x_{1:t}) \neq 0$. If we insert $\mathcal{X} = \{1,...,N\}$, $N = |\mathcal{X}|$, $i = x_t$, $y_i = \mu_t := \mu(x_t|x_{<t})$, and $z_i = \xi_t := \xi(x_t|x_{<t})$ into h and d we get (w.p.1)

$$h_t(x_{<t}) := \sum_{x_t}(\sqrt{\mu_t}-\sqrt{\xi_t})^2 \qquad \leq \qquad d_t(x_{<t}) := \sum_{x_t}\mu_t \ln\tfrac{\mu_t}{\xi_t} = \mathbf{E}_t[\ln\tfrac{\mu_t}{\xi_t}].$$

Taking the expectation $\mathbf{E}$ and the sum $\sum_{t=1}^n$ we get

$$\sum_{t=1}^{n}\mathbf{E}[d_t(x_{<t})] = \sum_{t=1}^{n}\mathbf{E}[\mathbf{E}_t[\ln\frac{\mu_t}{\xi_t}]] = \mathbf{E}[\ln\prod_{t=1}^{n}\frac{\mu_t}{\xi_t}] = \mathbf{E}[\ln\frac{\mu(x_{1:n})}{\xi(x_{1:n})}] \leq \ln w_\mu^{-1} \qquad (8)$$

where we have used $\mathbf{E}[\mathbf{E}_t[..]] = \mathbf{E}[..]$ and exchanged the t-sum with the expectation $\mathbf{E}$, which transforms to a product inside the logarithm. In the last equality we have used the chain rule for μ and ξ. Using universality $\xi(x_{1:n}) \geq w_\mu\mu(x_{1:n})$ yields the final inequality. Finally

$$\mathbf{E}_t\Big[\Big(\sqrt{\frac{\xi_t}{\mu_t}}-1\Big)^2\Big] = \sum_{x_t}{}'\mu_t\Big(\sqrt{\frac{\xi_t}{\mu_t}}-1\Big)^2 = \sum_{x_t}{}'(\sqrt{\xi_t}-\sqrt{\mu_t})^2 \leq h_t(x_{<t}) \leq d_t(x_{<t}).$$

Taking the expectation $\mathbf{E}$ and the sum $\sum_{t=1}^n$ and chaining the result with (8) yields Theorem 11. □

8 Convergence in Martin-Löf Sense

An interesting open question is whether ξ converges to μ (in difference or ratio) individually for all Martin-Löf random sequences. Clearly, 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 Solomonoff's universal prior M, since M.L. randomness can be defined in terms of M (see Theorem 8). Attempts to convert the bounds in Theorem 11 to effective μ.M.L.-randomness tests fail, since $M(x_t|x_{<t})$ is not enumerable. The proof of $M/\mu \stackrel{M.L.}{\longrightarrow} 1$ given in [LV97, Thm.5.2.2] and [VL00, Thm.10] is incomplete.[1] The implication "$M(x_{1:n}) \leq c\cdot\mu(x_{1:n})\forall n \Rightarrow \lim_{n\to\infty} M(x_{1:n})/\mu(x_{1:n})$ exists" has been used, but not proven, and is indeed generally wrong [HM04]. Theorem 8 only implies $\sup_n M(x_{1:n})/\mu(x_{1:n}) < \infty$ for M.L. random sequences $x_{1:\infty}$, and [Doo53, pp. 324–325] implies only that $\lim_{n\to\infty} M(x_{1:n})/\mu(x_{1:n})$ exists w.p.1, and not μ.M.L. Vovk [Vov87] shows that for two estimable semimeasures μ and ρ and $x_{1:\infty}$ being μ *and* ρ M.L. random that

$$\sum_{t=1}^{\infty}\sum_{x'_t}\left(\sqrt{\mu(x'_t|x_{<t})}-\sqrt{\rho(x'_t|x_{<t})}\right)^2<\infty \quad\text{and}\quad \sum_{t=1}^{\infty}\left(\frac{\rho(x_t|x_{<t})}{\mu(x_t|x_{<t})}-1\right)^2<\infty.$$

If M were estimable, then this would imply posterior $M\to\mu$ and $M/\mu\to 1$ for every μ.M.L.-random sequence $x_{1:\infty}$, since *every* sequence is M.M.L. random. Since M is *not* estimable, Vovk's theorem cannot be applied and it is not obvious how to generalize it. So the question of individual convergence remains open. More generally, one may ask whether $\xi_{\mathcal{M}}\to\mu$ for every μ/ξ-random sequence. It turns out that this is true for some $\mathcal{M}$, but false for others.

Theorem 12 (μ/ξ-convergence of ξ to μ) *Let $\mathcal{X}=\{0,1\}$ be binary and $\mathcal{M}_\Theta := \{\mu_\theta : \mu_\theta(1|x_{<t})=\theta\,\forall t,\ \theta\in\Theta\}$ be the set of Bernoulli(θ) distributions with parameters $\theta\in\Theta$. Let Θ_D be a countable dense subset of $[0,1]$, e.g. $[0,1]\cap\mathbb{Q}$, and let Θ_G be a countable subset of $[0,1]$ with a gap in the sense that there exist $0<\theta_0<\theta_1<1$ such that $[\theta_0,\theta_1]\cap\Theta_G=\{\theta_0,\theta_1\}$, e.g. $\Theta_G=\{\frac{1}{4},\frac{1}{2}\}$ or $\Theta_G=([0,\frac{1}{4}]\cup[\frac{1}{2},1])\cap\mathbb{Q}$. Then*

i) If $x_{1:\infty}$ is $\mu/\xi_{\mathcal{M}_{\Theta_D}}$ random with $\mu\in\mathcal{M}_{\Theta_D}$, then $\xi_{\mathcal{M}_{\Theta_D}}(x_t|x_{<t})\to\mu(x_t|x_{<t})$,

ii) There are $\mu\in\mathcal{M}_{\Theta_G}$ and $\mu/\xi_{\mathcal{M}_{\Theta_G}}$ random $x_{1:\infty}$ for which $\xi_{\mathcal{M}_{\Theta_G}}(x_t|x_{<t})\not\to\mu(x_t|x_{<t})$

[1] The formulation of their theorem is quite misleading in general: "*Let μ be a positive recursive measure. If the length of y is fixed and the length of x grows to infinity, then $M(y|x)/\mu(y|x)\to 1$ with μ-probability one. The infinite sequences ω with prefixes x satisfying the displayed asymptotics are precisely ['$\Rightarrow$' and '$\Leftarrow$'] the μ-random sequences.*" First, for off-sequence y convergence w.p.1 does not hold (xy must be demanded to be a prefix of ω). Second, the proof of '$\Leftarrow$' has gaps (see main text). Last, '$\Rightarrow$' is given without proof and is wrong [HM04]. Also the assertion in [LV97, Thm.5.2.1] that $S_t := \mathbf{E}\sum_{x'_t}(\mu(x'_t|x_{<t})-M(x'_t|x_{<t}))^2$ converges to zero faster than $1/t$ cannot be made, since S_t does not decrease monotonically [Hut04, Prob.2.7]. For example, for $a_t := 1/\sqrt{t}$ if t is a cube and 0 otherwise, we have $\sum_{t=1}^{\infty} a_t<\infty$, but $a_t\neq o(1/t)$.

Our original/main motivation of studying μ/ξ-randomness is the implication of Theorem 12 that $M \xrightarrow{\text{M.L.}} \mu$ cannot be decided from M being a mixture distribution or from the universality property (Theorem 3) alone. Further structural properties of $\mathcal{M}^{semi}_{enum}$ have to be employed. For Bernoulli sequences, convergence $\mu.\xi_{\mathcal{M}_\Theta}$.r. is related to denseness of $\mathcal{M}_\Theta$. Maybe a denseness characterization of $\mathcal{M}^{semi}_{enum}$ can solve the question of convergence M.L. of M. The property $M \in \mathcal{M}^{semi}_{enum}$ is also not sufficient to resolve this question, since there are $\mathcal{M} \ni \xi$ for which $\xi \xrightarrow{\mu.\xi.r} \mu$ and $\mathcal{M} \ni \xi$ for which $\xi \not\xrightarrow{\mu.\xi.r} \mu$. Theorem 12 can be generalized to i.i.d. sequences over general finite alphabet $\mathcal{X}$.

The idea to prove (ii) is to construct a sequence $x_{1:\infty}$ that is μ_{θ_0}/ξ-random *and* μ_{θ_1}/ξ-random for $\theta_0 \neq \theta_1$. This is possible if and only if Θ contains a gap and θ_0 and θ_1 are the boundaries of the gap. Obviously ξ cannot converge to θ_0 *and* θ_1, thus proving non-convergence. For no $\theta \in [0,1]$ will this $x_{1:\infty}$ be μ_θ M.L.-random. Finally, the proof of Theorem 12 makes essential use of the mixture representation of ξ, as opposed to the proof of Theorem 11 which only needs dominance $\xi \trianglerighteq \mathcal{M}$.

An example for (ii) is $\mathcal{M} = \{\mu_0, \mu_1\}$, $\mu_0(1|x_{<t}) = \mu_1(0|x_{<t}) = \frac{1}{4}$, $x_{1:\infty} = (01)^\infty = 01010101...$ $\Rightarrow$ $\mu_0(x_{1:2n}) = \mu_1(x_{1:2n}) = \xi(x_{1:2n}) = (\frac{1}{4})^n(\frac{3}{4})^n$ $\Rightarrow$ $x_{1:\infty}$ is μ_0/ξ-random *and* μ_1/ξ-random, but $\mu_0(x_{2n}|x_{<2n}) = \frac{1}{4}$, $\mu_0(x_{2n+1}|x_{1:2n}) = \frac{3}{4}$, $\mu_1(x_{2n}|x_{<2n}) = \frac{3}{4}$, $\mu_1(x_{2n+1}|x_{1:2n}) = \frac{1}{4}$ and $\xi(x_{2n}|x_{<2n}) = \frac{3}{8}$, $\xi(x_{2n+1}|x_{1:2n}) = \frac{1}{2}$ for $w_0 = w_1 = \frac{1}{2}$ $\Rightarrow$ $\xi(x_n|x_{<n}) \not\to \mu_{0/1}(x_n|x_{<n})$.

Proof. Let $\mathcal{X} = \{0,1\}$ and $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$ with countable $\Theta \subset [0,1]$ and $\mu_\theta(1|x_{1:n}) = \theta = 1 - \mu_\theta(0|x_{1:n})$, which implies

$$\mu_\theta(x_{1:n}) = \theta^{n_1}(1-\theta)^{n-n_1}, \qquad n_1 := x_1 + ... + x_n, \qquad \hat{\theta} \equiv \hat{\theta}_n := \frac{n_1}{n}$$

$\hat{\theta}$ depends on n; all other used/defined θ will be independent of n. We assume $\theta_{..} \in \Theta$, where .. stands for some (possible empty) index, and $\ddot{\theta} \in [0,1]$ (possibly $\notin \Theta$), where $\ddot{}$ stands for some superscript, i.e. $\mu_{\theta_{..}}$ and $w_{\theta_{..}}$ make sense, whereas $\mu_{\ddot{\theta}}$ and $w_{\ddot{\theta}}$ do not. ξ is defined in the standard way as

$$\xi(x_{1:n}) = \sum_{\theta \in \Theta} w_\theta \mu_\theta(x_{1:n}) \quad \Rightarrow \quad \xi(x_{1:n}) \geq w_\theta \mu_\theta(x_{1:n}), \tag{9}$$

where $\sum_\theta w_\theta = 1$ and $w_\theta > 0 \, \forall \theta$. In the following let $\mu = \mu_{\theta_0} \in \mathcal{M}$ be the true environment.

$$\omega = x_{1:\infty} \text{ is } \mu/\xi\text{-random} \quad \Leftrightarrow \quad \exists c_\omega : \xi(x_{1:n}) \leq c_\omega \cdot \mu_{\theta_0}(x_{1:n}) \, \forall n \tag{10}$$

For binary alphabet it is sufficient to establish whether $\xi(1|x_{1:n}) \overset{n\to\infty}{\longrightarrow} \theta_0 \equiv \mu(1|x_{1:n})$ for μ/ξ-random $x_{1:\infty}$ in order to decide $\xi(x_n|x_{<n}) \to \mu(x_n|x_{<n})$. We need the following posterior representation of ξ:

$$\xi(1|x_{1:n}) = \sum_{\theta \in \Theta} w_n^\theta \mu_\theta(1|x_{1:n}), \quad w_n^\theta := w_\theta \frac{\mu_\theta(x_{1:n})}{\xi(x_{1:n})} \leq \frac{w_\theta}{w_{\theta_0}} \frac{\mu_\theta(x_{1:n})}{\mu_{\theta_0}(x_{1:n})}, \quad \sum_{\theta \in \Theta} w_n^\theta = 1 \tag{11}$$

The ratio $\mu_\theta/\mu_{\theta_0}$ can be represented as follows:

$$\frac{\mu_\theta(x_{1:n})}{\mu_{\theta_0}(x_{1:n})} = \frac{\theta^{n_1}(1-\theta)^{n-n_1}}{\theta_0^{n_1}(1-\theta_0)^{n-n_1}} = \left[\left(\frac{\theta}{\theta_0}\right)^{\hat\theta_n}\left(\frac{1-\theta}{1-\theta_0}\right)^{1-\hat\theta_n}\right]^n = \mathrm{e}^{n[D(\hat\theta_n||\theta_0)-D(\hat\theta_n||\theta)]} \tag{12}$$

$$\text{where}\quad D(\hat\theta||\theta) = \hat\theta\ln\tfrac{\hat\theta}{\theta} + (1-\hat\theta)\ln\tfrac{1-\hat\theta}{1-\theta}$$

is the relative entropy between $\hat\theta$ and θ, which is continuous in $\hat\theta$ and θ, and is 0 if and only if $\hat\theta=\theta$. We also need the following implication for sets $\Omega\subseteq\Theta$:

$$\begin{aligned}&\text{If}\quad w_n^\theta \le w_\theta g_\theta(n) \stackrel{n\to\infty}{\longrightarrow} 0 \quad\text{and}\quad g_\theta(n)\le c\ \forall\theta\in\Omega,\\ &\text{then}\quad \sum_{\theta\in\Omega} w_n^\theta\mu_\theta(1|x_{1:n}) \le \sum_{\theta\in\Omega} w_n^\theta \stackrel{n\to\infty}{\longrightarrow} 0,\end{aligned} \tag{13}$$

which easily follows from boundedness $\sum_\theta w_n^\theta\le 1$ and $\mu_\theta\le 1$ [Hut04, Lem.5.28*ii*]. We now prove Theorem 12. We leave the special considerations necessary when $0,1\in\Theta$ to the reader and assume, henceforth, $0,1\notin\Theta$.

(i) Let Θ be a countable dense subset of (0,1) and $x_{1:\infty}$ be μ/ξ-random. Using (9) and (10) in (12) for $\theta\in\Theta$ to be determined later we can bound

$$\mathrm{e}^{n[D(\hat\theta_n||\theta_0)-D(\hat\theta_n||\theta)]} = \frac{\mu_\theta(x_{1:n})}{\mu_{\theta_0}(x_{1:n})} \le \frac{c_\omega}{w_\theta} =: c < \infty \tag{14}$$

Let us assume that $\hat\theta\equiv\hat\theta_n\not\to\theta_0$. This implies that there exists a cluster point $\tilde\theta\neq\theta_0$ of sequence $\hat\theta_n$, i.e. $\hat\theta_n$ is infinitely often in an ε-neighborhood of $\tilde\theta$, e.g. $D(\hat\theta_n||\tilde\theta)\le\varepsilon$ for infinitely many n. $\tilde\theta\in[0,1]$ may be outside Θ. Since $\tilde\theta\neq\theta_0$ this implies that $\hat\theta_n$ must be "far" away from θ_0 infinitely often. For instance, for $\varepsilon=\frac14(\tilde\theta-\theta_0)^2$, using $D(\hat\theta||\tilde\theta)+D(\hat\theta||\theta_0)\ge(\tilde\theta-\theta_0)^2$, we get $D(\hat\theta||\theta_0)\ge 3\varepsilon$. We now choose $\theta\in\Theta$ so near to $\tilde\theta$ such that $|D(\hat\theta||\theta)-D(\hat\theta||\tilde\theta)|\le\varepsilon$ (here we use denseness of Θ). Chaining all inequalities we get $D(\hat\theta||\theta_0)-D(\hat\theta||\theta)\ge 3\varepsilon-\varepsilon-\varepsilon=\varepsilon>0$. This, together with (14) implies $\mathrm{e}^{n\varepsilon}\le c$ for infinitely many n which is impossible. Hence, the assumption $\hat\theta_n\not\to\theta_0$ was wrong.

Now, $\hat\theta_n\to\theta_0$ implies that for arbitrary $\theta\neq\theta_0$, $\theta\in\Theta$ and for sufficiently large n there exists $\delta_\theta>0$ such that $D(\hat\theta_n||\theta)\ge 2\delta_\theta$ (since $D(\theta_0||\theta)\neq 0$) and $D(\hat\theta_n||\theta_0)\le\delta_\theta$. This implies

$$w_n^\theta \;\le\; \frac{w_\theta}{w_{\theta_0}}\mathrm{e}^{n[D(\hat\theta_n||\theta_0)-D(\hat\theta_n||\theta)]} \;\le\; \frac{w_\theta}{w_{\theta_0}}\mathrm{e}^{-n\delta_\theta} \;\stackrel{n\to\infty}{\longrightarrow}\; 0,$$

where we have used (11) and (12) in the first inequality and the second inequality holds for sufficiently large n. Hence $\sum_{\theta\neq\theta_0} w_n^\theta\to 0$ by (13) and $w_n^{\theta_0}\to 1$ by normalization (11), which finally gives

$$\xi(1|x_{1:n}) = w_n^{\theta_0}\mu_{\theta_0}(1|x_{1:n}) + \sum_{\theta\neq\theta_0} w_n^\theta\mu_\theta(1|x_{1:n}) \;\stackrel{n\to\infty}{\longrightarrow}\; \mu_{\theta_0}(1|x_{1:n}).$$

(ii) We first consider the case $\Theta = \{\theta_0,\theta_1\}$: Let us choose $\bar\theta$ $(= \ln(\frac{1-\theta_0}{1-\theta_1})/\ln(\frac{\theta_1}{\theta_0}\frac{1-\theta_0}{1-\theta_1}) \notin \Theta)$ in the (KL) middle of θ_0 and θ_1 such that

$$D(\bar\theta||\theta_0) = D(\bar\theta||\theta_1), \qquad 0 < \theta_0 < \bar\theta < \theta_1 < 1, \tag{15}$$

and choose $x_{1:\infty}$ such that $\hat\theta_n := \frac{n_1}{n}$ satisfies $|\hat\theta_n - \bar\theta| \leq \frac{1}{n}$ $(\Rightarrow\ \hat\theta_n \stackrel{n\to\infty}{\longrightarrow} \bar\theta)$

We will show that $x_{1:\infty}$ is μ_{θ_0}/ξ-random *and* μ_{θ_1}/ξ-random. Obviously no ξ can converge to θ_0 *and* θ_1, thus proving $\mathcal{M}$-non-convergence. ($x_{1:\infty}$ is obviously not $\mu_{\theta_{0/1}}$ M.L.-random, since the relative frequency $\hat\theta_n \not\to \theta_{0/1}$. $x_{1:\infty}$ is not even $\mu_{\bar\theta}$ M.L.-random, since $\hat\theta_n$ converges too fast ($\sim \frac{1}{n}$). $x_{1:\infty}$ is indeed very regular, whereas $\frac{n_1}{n}$ of a truly $\mu_{\bar\theta}$ M.L.-random sequence has fluctuations of the order $1/\sqrt{n}$. The fast convergence is necessary for doubly μ/ξ-randomness. The reason that $x_{1:\infty}$ is μ/ξ-random, but not M.L.-random is that μ/ξ-randomness is a weaker concept than M.L.-randomness for $\mathcal{M} \subset \mathcal{M}_{enum}^{semi}$. Only regularities characterized by $\nu \in \mathcal{M}$ are recognized by μ/ξ-randomness.)

In the following we assume that n is sufficiently large such that $\theta_0 \leq \hat\theta_n \leq \theta_1$. We need

$$|D(\hat\theta||\theta) - D(\bar\theta||\theta)| \leq c|\hat\theta - \bar\theta| \quad \forall\, \theta,\hat\theta,\bar\theta \in [\theta_0,\theta_1] \quad \text{with} \quad c := \ln\tfrac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)} < \infty \tag{16}$$

which follows for $\hat\theta \geq \bar\theta$ (similarly $\hat\theta \leq \bar\theta$) from

$$D(\hat\theta||\theta) - D(\bar\theta||\theta) = \int_{\bar\theta}^{\hat\theta} [\ln\tfrac{\theta'}{\theta} - \ln\tfrac{1-\theta'}{1-\theta}]d\theta' \leq \int_{\bar\theta}^{\hat\theta} [\ln\tfrac{\theta_1}{\theta_0} - \ln\tfrac{1-\theta_1}{1-\theta_0}]d\theta' = c\cdot(\hat\theta - \bar\theta)$$

where we have increased θ' to θ_1 and decreased θ to θ_0 in the inequality. Using (16) in (12) twice we get

$$\frac{\mu_{\theta_1}(x_{1:n})}{\mu_{\theta_0}(x_{1:n})} = \mathrm{e}^{n[D(\hat\theta_n||\theta_0) - D(\hat\theta_n||\theta_1)]} \leq \mathrm{e}^{n[D(\bar\theta||\theta_0) + c|\hat\theta_n - \bar\theta| - D(\bar\theta||\theta_1) + c|\hat\theta_n - \bar\theta|]} \leq \mathrm{e}^{2c} \tag{17}$$

where we have used (15) in the last inequality. Now, (17) and (11) lead to

$$w_n^{\theta_0} = w_{\theta_0}\frac{\mu_{\theta_0}(x_{1:n})}{\xi(x_{1:n})} = [1 + \frac{w_{\theta_1}}{w_{\theta_0}}\frac{\mu_{\theta_1}(x_{1:n})}{\mu_{\theta_0}(x_{1:n})}]^{-1} \geq [1 + \frac{w_{\theta_1}}{w_{\theta_0}}\mathrm{e}^{2c}]^{-1} =: c_0 > 0, \tag{18}$$

which shows that $x_{1:\infty}$ is μ_{θ_0}/ξ-random by (10). Exchanging $\theta_0 \leftrightarrow \theta_1$ in (17) and (18) we similarly get $w_n^{\theta_1} \geq c_1 > 0$, which implies (using $w_n^{\theta_0} + w_n^{\theta_1} = 1$)

$$\xi(1|x_{1:n}) = \sum_{\theta\in\{\theta_0,\theta_1\}} w_n^\theta \mu_\theta(1|x_{1:n}) = w_n^{\theta_0}\cdot\theta_0 + w_n^{\theta_1}\cdot\theta_1 \neq \theta_0 = \mu_{\theta_0}(1|x_{1:n}). \tag{19}$$

This shows $\xi(1|x_{1:n}) \stackrel{n\to\infty}{\not\longrightarrow} \mu(1|x_{1:n})$. One can show that $\xi(1|x_{1:n})$ does not only not converge to θ_0 (and θ_1), but that it does not converge at all. The fast convergence

demand $|\hat\theta_n-\bar\theta|\leq\frac{1}{n}$ on $x_{1:\infty}$ can be weakened to $\hat\theta_n\leq\bar\theta+O(\frac{1}{n})\,\forall n$ and $\hat\theta_n\geq\bar\theta-O(\frac{1}{n})$ for infinitely many n, then $x_{1:\infty}$ is still μ_{θ_0}/ξ-random, and $w_n^{\theta_1}\geq c_1'>0$ for infinitely many n, which is sufficient to prove $\xi\not\to\mu$.

We now consider general Θ with gap in the sense that there exist $0<\theta_0<\theta_1<1$ with $[\theta_0,\theta_1]\cap\Theta=\{\theta_0,\theta_1\}$: We show that all $\theta\neq\theta_0,\theta_1$ give asymptotically no contribution to $\xi(1|x_{1:n})$, i.e. (19) still applies. Let $\theta\in\Theta\setminus\{\theta_0,\theta_1\}$; all other definitions as before. Then $\delta_\theta:=D(\bar\theta||\theta)-D(\bar\theta||\theta_{0/1})>0$, since θ is farther than $\theta_{0/1}$ away from $\bar\theta$ ($|\theta-\bar\theta|>|\theta_{0/1}-\bar\theta|$). Similarly to (17) with θ instead θ_1 we get

$$\frac{\mu_\theta(x_{1:n})}{\mu_{\theta_0}(x_{1:n})}=\mathrm{e}^{n[D(\hat\theta_n||\theta_0)-D(\hat\theta_n||\theta)]}\leq\mathrm{e}^{2c}\cdot\mathrm{e}^{n[D(\bar\theta||\theta_0)-D(\bar\theta||\theta)]}=\mathrm{e}^{2c}\mathrm{e}^{-n\delta_\theta}\stackrel{n\to\infty}{\longrightarrow}0$$

Hence $w_n^\theta\leq\frac{w_\theta}{w_{\theta_0}}\mathrm{e}^{2c}\mathrm{e}^{-n\delta_\theta}\to0$ from (11) and $\varepsilon_n:=\sum_{\theta\in\Theta\setminus\{\theta_0,\theta_1\}}w_n^\theta\mu_\theta(1|x_{1:n})\stackrel{n\to\infty}{\longrightarrow}0$ from (13). Hence $\xi(1|x_{1:n})=w_n^{\theta_0}\cdot\theta_0+w_n^{\theta_1}\cdot\theta_1+\varepsilon_n\neq\theta_0=\mu_{\theta_0}(1|x_{1:n})$ for sufficiently large n, since $\varepsilon_n\to0$, $w_n^{\theta_1}\geq c_1'>0$ and $\theta_0\neq\theta_1$. □

9 Conclusions

For a hierarchy of four computability definitions, we completed the classification of the existence of computable (semi)measures dominating all computable (semi)measures. Dominance is an important property of a prior, since it implies rapid convergence of the corresponding posterior with probability one. A strengthening would be convergence for all Martin-Löf (M.L.) random sequences. This seems natural, since M.L. randomness can be defined in terms of Solomonoff's prior M, so there is a close connection. Contrary to what was believed before, the question of posterior convergence $M/\mu\to1$ for all M.L. random sequences is still open. Some exciting progress has been made recently in [HM04], partially answering this question. We introduced a new flexible notion of μ/ξ-randomness which contains Martin-Löf randomness as a special case. Though this notion may have a wider range of application, the main purpose for its introduction was to show that standard proof attempts of $M/\mu\stackrel{M.L.}{\longrightarrow}1$ based on dominance only must fail. This follows from the derived result that the validity of $\xi/\mu\to1$ for μ/ξ-random sequences depends on the Bayes mixture ξ.

References

[Cha75] G. J. Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM*, 22(3):329–340, 1975.

[Doo53] J. L. Doob. *Stochastic Processes*. Wiley, New York, 1953.

[Gác74] P. Gács. On the symmetry of algorithmic information. *Soviet Mathematics Doklady*, 15:1477–1480, 1974.

[HM04] M. Hutter and An. A. Muchnik. Universal convergence of semimeasures on individual random sequences. In *Proc. 15th International Conf. on Algorithmic Learning Theory (ALT-2004)*, volume 3244 of *LNAI*, pages 234–248, Padova, 2004. Springer, Berlin.

[Hut01] M. Hutter. Convergence and error bounds for universal prediction of nonbinary sequences. In *Proc. 12th European Conf. on Machine Learning (ECML-2001)*, volume 2167 of *LNAI*, pages 239–250, Freiburg, 2001. Springer, Berlin.

[Hut03a] 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, Sapporo, 2003. Springer, Berlin.

[Hut03b] M. Hutter. Sequence prediction based on monotone complexity. In *Proc. 16th Annual Conf. on Learning Theory (COLT-2003)*, volume 2777 of *LNAI*, pages 506–521, Washington, DC, 2003. Springer, Berlin.

[Hut04] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2004. 300 pages, http://www.idsia.ch/~marcus/ai/uaibook.htm.

[Kol65] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information and Transmission*, 1(1):1–7, 1965.

[Lam87] M. van Lambalgen. *Random Sequences.* PhD thesis, University of Amsterdam, 1987.

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

[Lev74] L. A. Levin. Laws of information conservation (non-growth) and aspects of the foundation of probability theory. *Problems of Information Transmission*, 10(3):206–210, 1974.

[LV97] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications.* Springer, Berlin, 2nd edition, 1997.

[Sch71] C. P. Schnorr. *Zufälligkeit und Wahrscheinlichkeit.* Springer, Berlin, 1971.

[Sch00] J. Schmidhuber. Algorithmic theories of everything. Report IDSIA-20-00, quant-ph/0011122, IDSIA, Manno (Lugano), Switzerland, 2000.

[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.

[Sim77] S. G. Simpson. Degrees of unsolvability: A survey of results. In J. Barwise, editor, *Handbook of Mathematical Logic*, pages 631–652. North-Holland, Amsterdam, 1977.

[Sol64] R. J. Solomonoff. A formal theory of inductive inference: Parts 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.

[Wan96] Y. Wang. *Randomness and Complexity*. PhD thesis, Universität Heidelberg, 1996.

[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.