On the Existence and Convergence of Computable Universal Priors*

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland

marcus@idsia.ch http://www.idsia.ch/~marcus

29 May 2003

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 his 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. We investigate the existence and convergence of computable universal (semi)measures for a hierarchy of computability classes: finitely computable, estimable, enumerable, and approximable. For instance, M is known to be enumerable, but not finitely computable, and to dominate all enumerable semimeasures. We define seven classes of (semi)measures based on these four computability concepts. Each class may or may not contain a (semi)measure which dominates all elements of another class. The analysis of these 49 cases can be reduced to four basic cases, two of them being new. The results hold for discrete and continuous semimeasures. We also 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.

Keywords

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

*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 and to a less extend 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. Solomonoff [Sol64, Sol78] formalized and combined both principles in his universal prior $M(x)$ which assigns high/low probability to simple/complex environments, hence implementing Occam and Epicurus. Solomonoff's [Sol78] central result is that if the probability $\mu(x_t|x_1...x_{t-1})$ of observing x_t at time t, given past observations $x_1...x_{t-1}$ is a computable function, then the universal posterior $M(x_t|x_1...x_{t-1})$ converges rapidly for $t\to\infty$ to the true posterior $\mu(x_t|x_1...x_{t-1})$, hence M represents a universal predictor in case of unknown μ.

One representation of M is as a weighted sum of *all* enumerable "defective" probability measures, called semimeasures (see Definition 2). The (from this representation obvious) dominance $M(x)\geq const.\times\mu(x)$ for all computable μ is the central ingredient in the convergence proof. What is so special about the class of all enumerable semimeasures $\mathcal{M}_{enum}^{semi}$? 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 finitely computable (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 does $\xi_\mathcal{M}$ possess. The distinguishing property of $M=\xi_{\mathcal{M}_{enum}^{semi}}$ is that it is itself an element of $\mathcal{M}_{enum}^{semi}$. 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 defined) to avoid getting into the (un)realm of non-constructive math.

The intention of this work is to investigate the existence, computability and convergence of universal (semi)measures for various computability classes: finitely computable $\subset$ estimable $\subset$ enumerable $\subset$ approximable (see Definition 1). For instance, $M(x)$ is enumerable, but not finitely computable. The research in this work was motivated by recent generalizations of Kolmogorov complexity and Solomonoff's prior by Schmidhuber [Sch02] to approximable (and others not here discussed) cases.

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 which 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 two new cases do not allow for domination. In Section 6 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 7 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 8 concludes the paper. Proofs will be presented elsewhere.

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, $l(x)$ for the length of string x, and $\omega=x_{1:\infty}$ for infinite sequences. 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 $f(x)\stackrel{\times}{\geq}g(x)$ for $g(x)=O(f(x))$.

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* finitely computable *or* recursive *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 *iff $\phi(\cdot,\cdot)$ is finitely computable and $\lim_{t\to\infty}\phi(x,t)=f(x)$.*
- *f is* lower semi-computable *or* enumerable *iff additionally $\phi(x,t)\leq\phi(x,t+1)$.*
- *f is* upper semi-computable *or* co-enumerable *iff $[-f]$ is lower semi-computable.*
- *f is* semi-computable *iff f is lower- or upper semi-computable.*
- *f is* estimable *iff f is lower- and upper semi-computable.*

If f is estimable we can finitely compute an ε-approximation of f by upper and lower semi-computing f and terminating when differing by less than ε. This means that there is a Turing machine which, given x and ε, finitely computes $\hat{y}$ 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 finitely computable (take any $\varepsilon<1$). Note that if f is only approximable or semi-computable we can still come arbitrarily close to $f(x)$ but we cannot devise a terminating algorithm which produces an ε-approximation. In the case of lower/upper semi-computability 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. In analogy to lower/upper semi-computability one may think of notions like lower/upper estimability but they are easily shown to coincide with estimability. The following implications are valid:

$$\begin{tikzcd} & & \boxed{\text{enumerable= lower semi-computable}} \arrow[dr, Rightarrow] & & \\ \boxed{\text{recursive= finitely computable}} \arrow[r, Rightarrow] & \boxed{\text{estimable}} \arrow[ur, Rightarrow] \arrow[dr, Rightarrow] & & \boxed{\text{semi-computable}} \arrow[r, Rightarrow] & \boxed{\text{approximable}} \\ & & \boxed{\text{co-enumerable= upper semi-computable}} \arrow[ur, Rightarrow] & & \end{tikzcd}$$

In the following we use the term computable synonymous to finitely computable, but sometimes also generically for some of the computability forms of Definition 1. What we call *estimable* is often just called *computable*, but it makes sense to separate the concepts of finite computability and estimability in this work, since the former is conceptually easier and some previous results have only been proved for this case.

3 The Universal Prior M

The prefix Kolmogorov complexity $K(x)$ is defined as the length of the shortest binary 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 [LV97]:

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

Solomonoff [Sol64, Sol78] (with a flaw fixed by Levin [ZL70]) defined (earlier) the closely related quantity, the universal prior $M(x)$. It is defined as the probability that the output of a universal Turing machine starts with x when provided with fair coin flips on the input tape. Formally, M can be defined as

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

where the sum is over all so called minimal programs p for which U outputs a string starting with x (indicated by the $*$). 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* $\mu(x)\geq\mu(x0)+\mu(x1)$*, and a (probability) measure if equality holds.*

We have $M(x0)+M(x1)<M(x)$ because there are programs p, which output x, neither followed by 0 nor 1. 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 [Sol78]:

Theorem 3 (Universality of M) *The universal prior M is an enumerable semimeasure which multiplicatively dominates all enumerable semimeasures in the sense that $M(x) \stackrel{\times}{\geq} 2^{-K(\rho)}\cdot\rho(x)$ for all an enumerable semimeasures ρ. M is enumerable, but not estimable or finitely computable.*

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 $2^{-K(x)}$ ($M(x)=2^{-K(x)+O(K(l(x)))}$), 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 (2) 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 [Hut03] for a more detailed analysis.

Sequence prediction algorithms try to predict the continuation $x_t\in\mathcal{X}$ of a given sequence $x_1...x_{t-1}$. We assume 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

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

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 9(i) and [Hut01] or Section 6 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 (2). 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

$$\xi(x) := \sum_{\nu\in\mathcal{M}} 2^{-K(\nu)}\nu(x). \qquad (3)$$

Then dominance

$$\xi(x) \geq 2^{-K(\nu)}\nu(x) \quad \forall\,\nu\in\mathcal{M} \qquad (4)$$

is obvious. Is ξ lower semi-computable? To answer this question one has to be more precise. Levin [ZL70] has shown that the set of *all* lower semi-computable semimeasures is enumerable (with repetitions). For this (ordered multi) set $\mathcal{M}=\mathcal{M}_{enum}^{semi}:=\{\nu_1,\nu_2,\nu_3,...\}$ and $K(\nu_i):=K(i)$ one can easily see that ξ is lower semi-computable. Finally proving $M(x)\stackrel{\times}{=}\xi(x)$ 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 $\mathcal{M}_{enum}^{semi}$? 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 finitely computable (semi)measures? It is clear that for every countable set $\mathcal{M}$,

$$\xi(x) := \xi_{\mathcal{M}}(x) := \sum_{\nu\in\mathcal{M}} w_\nu\nu(x) \quad\text{with}\quad \sum_{\nu\in\mathcal{M}} w_\nu \leq 1 \quad\text{and}\quad w_\nu>0 \qquad (5)$$

dominates all $\nu\in\mathcal{M}$. This dominance is necessary for the desired convergence $\xi\to\mu$ similarly to (2). The question is what properties ξ possesses. The distinguishing property of $\mathcal{M}_{enum}^{semi}$ is that ξ is itself an element of $\mathcal{M}_{enum}^{semi}$. 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

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

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

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

where $\mathcal{M}_c^{msr}$ stands for the set of all probability measures of appropriate computability type $c\in\{$comp=finitely computable, est=estimable, enum=enumerable, appr=approximable$\}$, and similarly for semimeasures $\mathcal{M}_c^{semi}$. From an enumeration of a measures ρ on 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 $\mathcal{M}_{enum}^{semi}\stackrel{\times}{\geq}\mathcal{M}_{enum}^{semi}$. Transitivity allows to conclude, for instance, that $\mathcal{M}_{appr}^{semi}\stackrel{\times}{\geq}\mathcal{M}_{comp}^{msr}$, i.e. that there is an approximable semimeasure which dominates all computable measures.

The standard "diagonalization" way of proving $\mathcal{M}_1\stackrel{\times}{\not\geq}\mathcal{M}_2$ is to take an arbitrary $\mu\in\mathcal{M}_1$ and "increase" it to ρ such that $\mu\stackrel{\times}{\not\geq}\rho$ 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 $\mathcal{M}_1\stackrel{\times}{\geq}\mathcal{M}_2$ 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) $\exists\rho:\{\rho\}\stackrel{\times}{\geq}\mathcal{M}$: For every countable set of (semi)measures $\mathcal{M}$, there is a (semi)measure which dominates all elements of $\mathcal{M}$.

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

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

iii) $\mathcal{M}_{est}^{semi} \overset{\times}{\not\geq} \mathcal{M}_{comp}^{msr}$: *There is* no *estimable semimeasure which dominates all computable measures.*

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

Table 5 (Existence of universal (semi)measures) *The entry in row r and column c indicates whether there is a r-able (semi)measure ρ for the set $\mathcal{M}$ which contains all c-able (semi)measures, where $r,c \in \{$comput, estimat, enumer, approxim$\}$. Enumerable measures are estimable. This is the reason why the enum. row and column in case of measures is 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 $\overset{\times}{\geq}$.*

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

If we ask for a universal (semi)measure which 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 $\mathcal{M}_{enum}^{semi}$ plays a special role. On the other hand, $\mathcal{M}_{enum}^{semi}$ 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 $\mathcal{M}_{CEM}$, has recently been discovered by Schmidhuber [Sch02], for which even $\xi_{CEM} \in \mathcal{M}_{CEM}$ holds.
Theorem 4 also holds for *discrete (semi)measures P* defined as follows:

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

Theorem 4 (*i*) is Levin's major result [LV97, Th.4.3.1 & Th.4.5.1], (*ii*) is due to Solomonoff [Sol78], the proof of $\mathcal{M}_{comp}^{semi} \overset{\times}{\not\geq} \mathcal{M}_{comp}^{semi}$ 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. A complete proof of *(o)*$-$*(iv)* for discrete and continuous (semi)measures is given elsewhere.

6 Posterior Convergence

We have investigated in detail the computational properties of various mixture distributions ξ. A mixture $\xi_{\mathcal{M}}$ multiplicatively dominates all distributions in $\mathcal{M}$. We have 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 and "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 7 (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.*

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 8 (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 (3), in which case the reverse inequality $\xi(x)\stackrel{\times}{\geq}\mu(x)$ 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 used in ξ. For $\mathcal{M}=\mathcal{M}_{enum}^{semi}$, μ/ξ-randomness is just μ.M.L. randomness. The larger $\mathcal{M}$, the more patterns are recognized as non-random. Roughly speaking, those regularities characterized by some $\nu\in\mathcal{M}$ are recognized by μ/ξ-randomness, i.e. for $\mathcal{M}\subset\mathcal{M}_{enum}^{semi}$ 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 (non-random) 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 $f:\mathcal{X}^t\to I\!R$, 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 9 (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 $z_*(\omega)$*

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.) $:\Leftrightarrow$
$\forall\omega$: $[\exists c\forall n:M(\omega_{1:n})\leq c\mu(\omega_{1:n})]$ *implies* $z_t(\omega)\to z_*(\omega)$ *for* $t\to\infty$,

iv) for every μ/ξ-random sequence ($\mu.\xi$.r.) $:\Leftrightarrow$
$\forall\omega$: $[\exists c\forall n:\xi(\omega_{1:n})\leq c\mu(\omega_{1:n})]$ *implies* $z_t(\omega)\to z_*(\omega)$ *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 $\sum_{t=1}^\infty\mathbf{E}[(z_t-z_*)^2]/\varepsilon^2$. 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 10 (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. The universal 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:*

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

where w_μ is the weight (5) of μ in ξ.

Theorem 10 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 in [LV97, Th.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* the μ-random sequence ($x'_{1:\infty}=x_{1:\infty}$), we have $\xi(x_t|x_{<t})/\mu(x_t|x_{<t})\to 1$ (whether $\inf\mu(x_t|x_{<t})$ tends to zero or not does not matter). Indeed, it is easy to see that $\xi(1|0_{<t})/\mu(1|0_{<t})\propto t\to\infty$ diverges for $\mathcal{M}=\{\mu,\nu\}$, $\mu(1|x_{<t}):=\frac{1}{2}t^{-3}$ and $\nu(1|x_{<t}):=\frac{1}{2}t^{-2}$, although $0_{1:\infty}$ is μ-random.

7 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 7). Attempts to convert the bounds in Theorem 10 to effective μ.M.L. randomness tests fail, since $M(x_t|x_{<t})$ is not enumerable. The proof given of $M/\mu\stackrel{M.L.}{\longrightarrow}1$ in [LV97, Th.5.2.2] and [VL00, Th.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 may indeed be wrong.

Vovk [Vov87] shows that for two finitely computable semi-measures 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 recursive, 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* recursive Vovk's theorem cannot be applied and it is not obvious how

[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 '$\Leftarrow$' is loopy (see main text). Last, '$\Rightarrow$' is given without proof and is probably wrong. Also the assertion in [LV97, Th.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 may not decrease monotonically.

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 11 (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 $\xi_{\mathcal{M}_{\Theta_G}}(x_t|x_{<t})\not\to\mu(x_t|x_{<t})$

Our original/main motivation of studying μ/ξ-randomness is the implication of Theorem 11 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 $\mathcal{M}_{enum}^{semi}$ 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 $\mathcal{M}_{enum}^{semi}$ can solve the question of convergence M.L. of M. The property $M\in\mathcal{M}_{enum}^{semi}$ 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 $\xi \overset{\mu.\xi.r}{\not\longrightarrow} \mu$. Theorem 11 can be generalized to i.i.d. sequences over general finite alphabet $\mathcal{X}$.

The idea to prove (ii) is to construct a sequence $x_{1:\infty}$ which is $\mu_{\theta_0}\mathcal{M}$-random *and* $\mu_{\theta_1}\mathcal{M}$-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 $\mathcal{M}$-non-convergence. For no $\theta\in[0,1]$ will this $x_{1:\infty}$ be μ_θ M.L.-random. Finally, the proof of Theorem 11 makes essential use of the mixture representation of ξ, as opposed to the proof of Theorem 10 which only needs dominance $\xi \overset{\times}{\geq} \mathcal{M}$.

8 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 $M/\mu\to 1$ for all M.L. random sequences is still open. 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 $\xi/\mu \to 1$ for μ/ξ-random sequences depends on the Bayes mixture ξ.

References

[Hut01] M. Hutter. Convergence and error bounds of universal prediction for general alphabet. *Proceedings of the 12th Eurpean Conference on Machine Learning (ECML-2001)*, pages 239–250, 2001.

[Hut03] M. Hutter. Sequence prediction based on monotone complexity. Technical Report IDSIA-09-03, 2003.

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

[Lev73] L. A. Levin. On the notion of a random sequence. *Soviet Math. Dokl.*, 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.

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

[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. *Inform. Control*, 7:1–22, 224–254, 1964.

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

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

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

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