

# UNIVERSAL CONVERGENCE OF SEMIMEASURES ON INDIVIDUAL RANDOM SEQUENCES

---

Marcus Hutter

IDSIA, Galleria 2

CH-6928 Manno-Lugano

Switzerland

<http://www.idsia.ch/~marcus>

Andrej Muchnik

Institute of New Technologies

10 Nizhnyaya Radischewskaya

Moscow 109004, Russia

[muchnik@lpcs.math.msu.ru](mailto:muchnik@lpcs.math.msu.ru)

ALT-2004, October 2-5

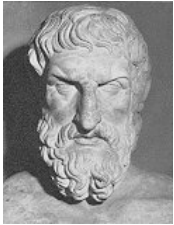
# Abstract

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

# Table of Contents

- Induction = Predicting the Future
- Meaning of Randomness and Probability
- Solomonoff's Universal Prior  $M$
- (Semi)measures & Universality
- Martin-Löf Randomness
- Convergence of Random Sequences
- Posterior Convergence  $M$
- Failed Attempts to Prove  $M \xrightarrow{\text{M.L.}} \mu$
- Non-M.L.-Convergence of  $M$
- M.L.-Converging Enumerable Semimeasure  $W$
- Open Problems

# Induction = Predicting the Future



## Epicurus' principle of multiple explanations

If more than one theory is consistent with the observations, keep all theories.



## Ockhams' razor (simplicity) principle

Entities should not be multiplied beyond necessity.



## Hume's negation of Induction

The only form of induction possible is deduction as the conclusion is already logically contained in the start configuration.



## Bayes' rule for conditional probabilities

Given the prior believe/probability one can predict all future probabilities.



## Solomonoff's universal prior

Solves the question of how to choose the prior if nothing is known.

# When is a Sequence Random?

- a) Is 0110010100101101101001111011 generated by a fair coin flip?
- b) Is 11111111111111111111111111111111 generated by a fair coin flip?
- c) Is 1100100100001111110110101010 generated by a fair coin flip?
- d) Is 01010101010101010101010101010101 generated by a fair coin flip?

- Intuitively: (a) and (c) look random, but (b) and (d) look unlikely.
- Problem: Formally (a-d) have equal probability  $(\frac{1}{2})^{length}$ .
- Classical solution: Consider hypothesis class  $H := \{\text{Bernoulli}(p) : p \in \Theta \subseteq [0, 1]\}$  and determine  $p$  for which sequence has maximum likelihood  $\implies$  (a,c,d) are fair Bernoulli( $\frac{1}{2}$ ) coins, (b) not.
- Problem: (d) is non-random, also (c) is binary expansion of  $\pi$ .
- Solution: Choose  $H$  larger, but how large? Overfitting? MDL?
- AIT Solution: A sequence is **random** iff it is **incompressible**.

# What does Probability Mean?

Naive frequency interpretation is circular:

- Probability of event  $E$  is  $p := \lim_{n \rightarrow \infty} \frac{k_n(E)}{n}$ ,  
 $n = \#$  i.i.d. trials,  $k_n(E) = \#$  occurrences of event  $E$  in  $n$  trials.
- Problem: Limit may be anything (or nothing):  
e.g. a fair coin can give: Head, Head, Head, Head, ...  $\Rightarrow p = 1$ .
- Of course, for a fair coin this sequence is “unlikely”.  
For fair coin,  $p = 1/2$  with “high probability”.
- But to make this statement rigorous we need to formally know what  
“high probability” means. **Circularity!**

Also: In complex domains typical for AI, sample size is often 1.

(e.g. a single non-iid historic weather data sequences is given).

We want to know whether certain properties hold for this *particular* seq.

# Solomonoff's Universal Prior $M$

**Strings:**  $x = x_1x_2\dots x_n$  with  $x_t \in \{0, 1\}$  and  $x_{1:n} := x_1x_2\dots x_{n-1}x_n$  and  $x_{<n} := x_1\dots x_{n-1}$ .

**Probabilities:**  $\rho(x_1\dots x_n)$  is the probability that an (infinite) sequence starts with  $x_1\dots x_n$ .

**Conditional probability:**  $\rho(x_t|x_{<t}) = \rho(x_{1:t})/\rho(x_{<t})$  is the  $\rho$ -probability that a given string  $x_1\dots x_{t-1}$  is followed by (continued with)  $x_t$ .

The **universal prior**  $M(x)$  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

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

## Semimeasures & Universality

Continuous (Semi)measures:  $\mu(x) \stackrel{(\geq)}{=} \mu(x0) + \mu(x1)$  and  $\mu(\epsilon) \stackrel{(\leq)}{=} 1$ .  
 $\mu(x)$  = probability that a sequence starts with string  $x$ .

Universality of  $M$  (Levin:70):  $M$  is an enumerable semimeasure.  
 $M(x) \geq w_\rho \cdot \rho(x)$  with  $w_\rho = 2^{-K(\rho) - O(1)}$  for all enum. semimeas.  $\rho$ .

Explanation: Up to a multiplicative constant,  $M$  assigns higher probability to all  $x$  than any other computable probability distribution.



# Martin-Löf Randomness

- Martin-Löf randomness is a very important concept of randomness of individual sequences.
- Characterization by Levin:73: Sequence  $x_{1:\infty}$  is  $\mu$ -Martin-Löf random ( $\mu$ .M.L.)  $\Leftrightarrow \exists c : M(x_{1:n}) \leq c \cdot \mu(x_{1:n}) \forall n$ . Moreover,  $d_\mu(\omega) := \sup_n \left\{ \log \frac{M(\omega_{1:n})}{\mu(\omega_{1:n})} \right\} \leq \log c$  is called the randomness deficiency of  $\omega := x_{1:\infty}$ .
- A  $\mu$ .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  $\mu$ .M.L. random sequences has  $\mu$ -measure 1.

# Convergence of Random Sequences

Let  $z_1(\omega), z_2(\omega), \dots$  be a sequence of real-valued random variables.

$z_t$  is said to converge for  $t \rightarrow \infty$  to random variable  $z_*(\omega)$

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

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

*iii)* for every  $\mu$ -Martin-Löf random sequence ( **$\mu$ .M.L.**)  $:\Leftrightarrow$

$\forall \omega : [\exists c \forall n : M(\omega_{1:n}) \leq c \cdot \mu(\omega_{1:n})]$  implies  $z_t(\omega) \xrightarrow{t \rightarrow \infty} z_*(\omega),$

where  $\mathbf{E}[\dots]$  denotes the expectation and  $\mathbf{P}[\dots]$  denotes the probability of  $[\dots]$ .

## Remarks

(i) In statistics, convergence **w.p.1** is the “**default**” characterization of convergence of random sequences.

(ii) Convergence **i.m.s.** 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  $\varepsilon$  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.

(iii) **Martin-Löf's** notion of randomness of **individual** sequences.

Convergence i.m.s. implies convergence w.p.1 + convergence rate.

Convergence M.L. implies convergence w.p.1 + on which sequences.

# Posterior Convergence

**Theorem:** Universality  $M(x) \geq w_\mu \mu(x)$  implies the following posterior convergence results for the Hellinger distance

$$h_t(M, \mu | \omega_{<t}) := \sum_{a \in \mathcal{X}} (\sqrt{M(a | \omega_{<t})} - \sqrt{\mu(a | \omega_{<t})})^2$$

$$\sum_{t=1}^{\infty} \mathbf{E} \left[ \left( \sqrt{\frac{M(\omega_t | \omega_{<t})}{\mu(\omega_t | \omega_{<t})}} - 1 \right)^2 \right] \leq \sum_{t=1}^{\infty} \mathbf{E}[h_t] \leq 2 \ln \{ \mathbf{E}[\exp(\frac{1}{2} \sum_{t=1}^{\infty} h_t)] \} \leq \ln w_\mu^{-1}$$

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

**Implications:**

$$M(x'_t | x_{<t}) \rightarrow \mu(x'_t | x_{<t}) \quad \text{for any } x'_t \text{ rapid w.p.1 for } t \rightarrow \infty.$$

$$\frac{M(x_t | x_{<t})}{\mu(x_t | x_{<t})} \rightarrow 1 \quad \text{rapid w.p.1 for } t \rightarrow \infty.$$

The probability that the number of  $\varepsilon$ -deviations of  $M_t$  from  $\mu_t$  exceeds  $\frac{1}{\varepsilon^2} \ln w_\mu^{-1}$  is small.

**Question:** Does  $M_t$  converge to  $\mu_t$  for all Martin-Löf random sequences?

## Failed Attempts to Prove $M \xrightarrow{\text{M.L.}} \mu$

- Conversion of bound to effective  $\mu$ .M.L. randomness tests fails, since they are not enumerable.
- The proof given in Vitanyi&Li:00 is erroneous.
- Vovk:87 shows that for two finitely computable (semi)measures  $\mu$  and  $\rho$  and  $x_{1:\infty}$  being  $\mu$ .M.L. random that
$$\sum_{t=1}^{\infty} \left( \sqrt{\mu(x_t|x_{<t})} - \sqrt{\rho(x_t|x_{<t})} \right)^2 < \infty \Leftrightarrow x_{1:\infty} \text{ is } \rho\text{-M.L. random.}$$
If  $M$  were recursive, then this would imply  $M \rightarrow \mu$  for every  $\mu$ .M.L. random sequence  $x_{1:\infty}$ , since *every* sequence is  $M$ .M.L. random.
- $M \xrightarrow{\text{M.L.}} \mu$  cannot be decided from  $M$  being a mixture distribution or from dominance or enumerability alone.

# Universal Semimeasure (USM) Non-Convergence

- $M \not\rightarrow \mu$ : There exists a universal semimeasure  $M$  and a computable measure  $\mu$  and a  $\mu$ .M.L.-random sequence  $\alpha$ , such that

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

- **Proof idea:** construct  $\nu$  such that  $\nu$  dominates  $M$  on some  $\mu$ -random sequence  $\alpha$ , but  $\nu(\alpha_n | \alpha_{<n}) \not\rightarrow \mu(\alpha_n | \alpha_{<n})$ . Then pollute  $M$  with  $\nu$ .
- **Open problem:** There may be *some* universal semimeasures for which convergence holds.
- **Converse:**  $\forall$  USM  $M \exists$  comp.  $\mu$  and non- $\mu$ .M.L.-random sequences  $\alpha$  for which  $M(\alpha_n | \alpha_{<n}) / \mu(\alpha_n | \alpha_{<n}) \rightarrow 1$ .

# Convergence in Martin-Löf Sense

**Main result:** There exists a *non-universal* enumerable semimeasure  $W$  such that for every computable measure  $\mu$  and every  $\mu$ .M.L.-random sequence  $\omega$ , the posteriors converge to each other:

$$W(a|\omega_{<t}) \xrightarrow{t \rightarrow \infty} \mu(a|\omega_{<t}) \quad \text{for all } a \in \mathcal{X} \quad \text{if } d_\mu(\omega) < \infty.$$

We need a converse of “M.L. implies w.p.1”.

## Lemma: Conversion of Expected Bounds to Individual Bounds

Let  $F(\omega) \geq 0$  be an enumerable function and  $\mu$  be an enumerable measure and  $\varepsilon > 0$  be co-enumerable. Then:

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

$K(\mu, F, 1/\varepsilon)$  is the complexity of  $\mu$ ,  $F$ , and  $1/\varepsilon$ .

**Proof:** Integral test  $\rightarrow$  submartingale  $\rightarrow$  universal submartingale  $\rightarrow$  rand.defect

# Convergence in Martin-Löf Sense of $D$

Mixture over proper computable measures:

$$J_k := \{i \leq k : \nu_i \text{ is measure}\}, \quad \varepsilon_i = i^{-6} 2^{-i}$$

$$\delta_k(x) := \sum_{i \in J_k}^{\times} \varepsilon_i \nu_i(x), \quad D(x) := \delta_{\infty}(x)$$

**Theorem:** If  $\mu = \nu_{k_0}$  is a computable measure, then

$$i) \quad \sum_{t=1}^{\infty} h_t(\delta_{k_0}, \mu) \stackrel{+}{<} 2 \ln 2 \cdot d_{\mu}(\omega) + 3k_0$$

$$ii) \quad \sum_{t=1}^{\infty} h_t(\delta_{k_0}, D) \stackrel{\times}{<} k_0^7 2^{k_0 + d_{\mu}(\omega)}$$

Although  $J_k$  and  $\delta_k$  are non-constructive, they are computable!

But  $J_{\infty}$  and  $D$  are *not* computable, not even approximable :-)



# M.L.-Converging Enumerable Semimeasure $W$

**Idea:** Enlarge the class of computable measures to an enumerable class of semimeasures, which are still sufficiently close to measures in order not to spoil the convergence result.

**Quasimeasures:**  $\tilde{\nu}(x_{1:n}) := \nu(x_{1:n})$  if  $\sum_{y_{1:n}} \nu(y_{1:n}) > 1 - \frac{1}{n}$ , and 0 else.

**Enumerable semimeasure:**  $W(x) := \sum_{i=1}^{\infty} \varepsilon_i \tilde{\nu}_i(x) \implies$

**Theorem:**  $\frac{W(\omega_{1:t})}{D(\omega_{1:t})} \rightarrow 1, \quad \frac{W(\omega_t|\omega_{<t})}{D(\omega_t|\omega_{<t})} \rightarrow 1, \quad W(a|\omega_{<t}) \rightarrow D(a|\omega_{<t})$

**Proof:** Additional contributions of non-measures to  $W$  absent in  $D$  are zero for long sequences.

**Together:**  $W \rightarrow D$  and  $D \rightarrow \mu \implies W \rightarrow \mu.$

# Properties of $M$ , $D$ , and $W$ – Summary

- $M :=$  mixture-over-semimeasures is an enumerable semimeasure, which dominates all enumerable semimeasures.  $M$  is not computable and not a measure.

$M \xrightarrow[\text{fast}]{w.p.1} \mu$  with logarithmic bound in  $w_\mu$ , but  $M \not\xrightarrow{M.L.} \mu$ .

- $D :=$  mixture-over-measures is a measure, dominating all enumerable quasimeasures.  $D$  is not computable and does not dominate all enumerable semimeasures.

$D \xrightarrow{M.L.} \mu$ , but bound is exponential in  $w_\mu$ .

- $W :=$  mixture-over-quasimeasures is an enumerable semimeasure, which dominates all enumerable quasimeasures.  $W$  is not itself a quasimeasure, is not computable, and does not dominate all enumerable semimeasures.

$W \xrightarrow{M.L.} \mu$ , asymptotically (bound exponential in  $w_\mu$  can be shown).

## Open Problems

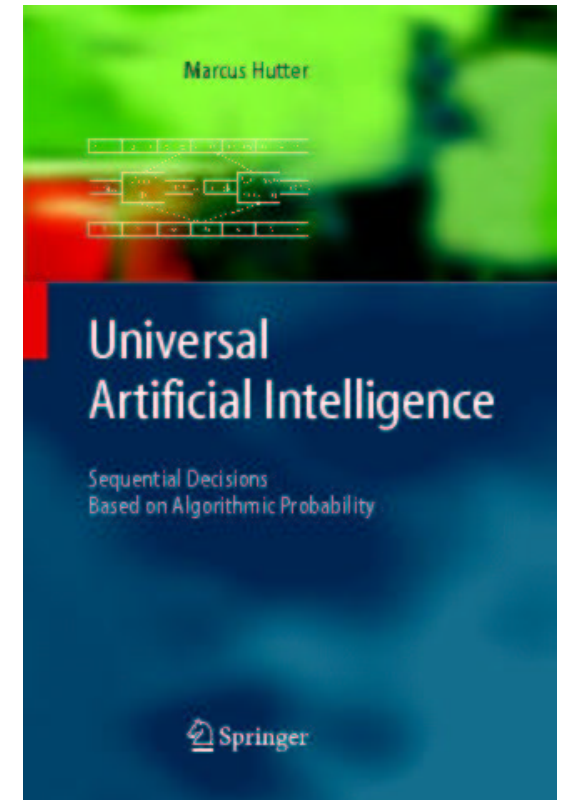
- The bounds for  $D \xrightarrow{M.L.} \mu$  and  $W \xrightarrow{M.L.} D$  are double exponentially worse than for  $M \xrightarrow{w.p.1} \mu$ . Can this be improved?
- Finally there could still exist **universal** semimeasures  $M$  (dominating all enumerable semimeasures) for which **M.L.-convergence** holds ( $\exists M : M \xrightarrow{M.L.} \mu ?$ ).
- In case they exist, we expect them to have particularly interesting **additional structure and properties**.
- Identify a class of “**natural**” **UTMs/USMs** which have a variety of favorable properties.

See [www.idsia.ch/~marcus/ai/uaibook.htm](http://www.idsia.ch/~marcus/ai/uaibook.htm) for prizes.

# Thanks! Questions? Details:

**Papers** at <http://www.idsia.ch/~marcus>

**Book** intends to excite a broader AI audience about abstract Algorithmic Information Theory –and– inform theorists about exciting applications to AI.



$$\begin{array}{rcl}
 \text{Decision Theory} & = & \text{Probability} + \text{Utility Theory} \\
 + & & + \\
 \text{Universal Induction} & = & \text{Ockham} + \text{Bayes} + \text{Turing} \\
 = & & = \\
 \text{A Unified View of Artificial Intelligence} & & 
 \end{array}$$

**Open research problems** at [www.idsia.ch/~marcus/ai/uaibook.htm](http://www.idsia.ch/~marcus/ai/uaibook.htm)