On the Existence and Convergence of Computable Universal Priors

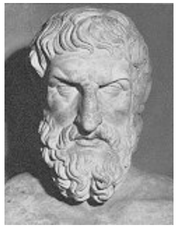
Marcus Hutter

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch,   http://www.idsia.ch/~marcus

2003

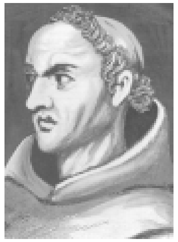
Table of Contents

Induction = Predicting the Future



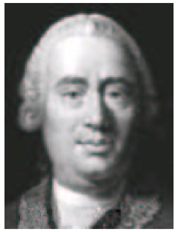
Epicurus' principle of multiple explanations

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



Ockhams' razor (simplicity) principle

Entities should not be multiplied beyond necessity.

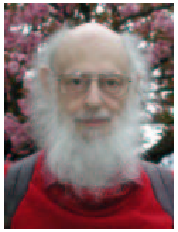


Hume's negation of Induction

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.

Strings and Conditional Probabilities

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

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

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

Interpretation of Probabilities

Frequentist: Probabilities come from experiments.

Objectivist: Probabilities are real aspects of the world.

Subjectivist: Probabilities describe ones believe.

Computability Concepts

f is finitely computable or recursive *iff* there are Turing machines $T_{1/2}$ with output interpreted as natural numbers and $f(x) = \frac{T_1(x)}{T_2(x)}$,

$$\Downarrow$$

f is estimable *iff* $\exists$ recursive $\phi(\cdot,\cdot)$ $\forall\, \varepsilon>0$: $|\phi(x,\lfloor\frac{1}{\varepsilon}\rfloor) - f(x)| < \varepsilon$ $\forall x$.

$$\Downarrow$$

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

$$\Downarrow$$

f is approximable *iff* $\phi(\cdot,\cdot)$ is recursive and $\lim_{t\to\infty}\phi(x,t) = f(x)$.

(What we call estimable is often just called computable)

Kolmogorov Complexity & Solomonoff Prior

The prefix Kolmogorov complexity of a string x is the length of the shortest (prefix) program p (on a universal Turing machine U) producing x (given y)

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

Solomonoff:64 (with a flaw fixed by Levin:70) defined (earlier) the closely related universal prior $M(x)$

$M(x)$ 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, Normalization

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

Universality of M (Solomonoff:78): M is an enumerable semimeasure.
$M(x) \geq w_\rho \cdot \rho(x)$ with $w_\rho = 2^{-K(\rho)-O(1)}$ for all an enum. semimeas. ρ.

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

Normalization: It is possible to normalize M to a true probability measure M_{norm} with dominance still being true, but at the expense of giving up enumerability (M_{norm} is still approximable).

Bayes-Mixtures and Dominance

Consider a countable set of semimeasures $\mathcal{M}$, $w_\nu > 0$, $\xi = \xi_\mathcal{M}$:

$$\xi(x) := \sum_{\nu\in\mathcal{M}} w_\nu \nu(x) \quad\Rightarrow\quad \xi(x) \geq w_\nu \nu(x) \quad\Rightarrow\quad \xi(x_t|x_{<t}) \to \nu(x_t|x_{<t})$$

Example: $\mathcal{M} = \mathcal{M}_{enum}^{semi} = \{enumerable\ semimeasures\} \;\Rightarrow\; \xi \stackrel{\times}{=} M$.

The distinguishing property of $\mathcal{M}_{enum}^{semi}$ is that $\xi \in \mathcal{M}_{enum}^{semi}$.

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 defined senses.

$$\mathcal{M}_1 \stackrel{\times}{\geq} \mathcal{M}_2 :\Leftrightarrow \exists\rho \in \mathcal{M}_1\, \forall\nu \in \mathcal{M}_2\, \exists w_\nu > 0\, \forall x : \rho(x) \geq w_\nu \nu(x).$$

$\stackrel{\times}{\geq}$ is transitive (but not necessarily reflexive) in the sense that

$$\mathcal{M}_1 \stackrel{\times}{\geq} \mathcal{M}_2 \stackrel{\times}{\geq} \mathcal{M}_3 \Rightarrow \mathcal{M}_1 \stackrel{\times}{\geq} \mathcal{M}_3 \text{ and } \mathcal{M}_0 \supseteq \mathcal{M}_1 \stackrel{\times}{\geq} \mathcal{M}_2 \supseteq \mathcal{M}_3 \Rightarrow \mathcal{M}_0 \stackrel{\times}{\geq} \mathcal{M}_3$$

(Semi)Computable (Semi)Measures

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

- With this notation, Levin's result reads: $\mathcal{M}^{semi}_{enum} \stackrel{\times}{\geq} \mathcal{M}^{semi}_{enum}$.
- 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 wrong.
- The 49 combinations follow by transitivity from 4 basic cases:

Universal (Semi)Measures

A semimeasure ρ is universal for $\mathcal{M}$ if it multiplicatively dominates all elements of $\mathcal{M}$ in the sense $\forall\nu\exists w_\nu>0:\rho(x)\geq w_\nu\nu(x)\forall x$:

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

i) $\mathcal{M}_{enum}^{semi}\overset{\times}{\geq}\mathcal{M}_{enum}^{semi}$: 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.

Universal (Semi)Measures

$\nwarrow$	$\mathcal{M}$	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^{iii}	no^{iii}	no^{iv}
m	enum.	yes^{i}	yes^{i}	YES^{i}	no^{iv}	yes^{i}	yes^{i}	no^{iv}
i	appr.	yes^{i}	yes^{i}	yes^{i}	no^{iv}	yes^{i}	yes^{i}	NO^{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^{ii}	no^{iv}	yes^{ii}	yes^{ii}	no^{iv}

Discussion

- 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:02, for which even $\xi_{CEM} \in \mathcal{M}_{CEM}$ holds.
- The dominance properties also holds for discrete (semi)measures $P : I\!N \to [0,1]$ with $\sum_{x\in I\!N} P(x) \overset{(\leq)}{=} 1$.

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 μ-Martin-Löf random (μ.M.L.) $\Leftrightarrow \exists c : M(x_{1:n}) \leq c \cdot \mu(x_{1:n}) \forall n$.

- 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. Especially, the set of all μ.M.L. random sequences has μ-measure 1.

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

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

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

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

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

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 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.

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

(iv) μ/ξ-randomness based on ξ **generalizes** the definition of M.L. randomness based on M.

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

Convergence M.L. implies convergence w.p.1.

Posterior Convergence

Universality $\xi(x) \geq w_\mu \mu(x)$ implies the following posterior convergence results:

$$i) \quad \sum_{t=1}^{n} \mathbf{E} \sum_{x'_t} \big(\mu(x'_t|x_{<t}) - \xi(x'_t|x_{<t})\big)^2 \;\leq\; \ln w_\mu^{-1} \;<\; \infty$$

$\xi(x'_t|x_{<t}) \to \mu(x'_t|x_{<t})$ for any x'_t i.m.s. for $t \to \infty$.

$$ii) \quad \sum_{t=1}^{n} \mathbf{E} \left[\left(\sqrt{\frac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})}} - 1\right)^2\right] \;\leq\; \ln w_\mu^{-1} \;<\; \infty$$

$\sqrt{\frac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})}} \to 1$ i.m.s. for $t \to \infty$.

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.

Failed Attempts to Proof $M \xrightarrow{\text{M.L.}} \mu$:

- Conversion of bounds (i) or (ii) to effective μ.M.L. randomness tests fails, since they are not enumerable.

- The proof given in Vitanyi&Li:00 is incomplete. 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:87 shows that for two finitely computable (semi)measures μ and ρ and $x_{1:\infty}$ being μ.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 $M\to\mu$ for every μ.M.L. random sequence $x_{1:\infty}$, since every sequence is M.M.L. random.

Generalization

- More generally, one may ask whether $\xi \to \mu$ for every μ/ξ-random sequence.
- It turns out that this is true for some $\mathcal{M}$, but wrong for others.
- This implies that $M \xrightarrow{\text{M.L.}} \mu$ cannot be decided from M being a mixture distribution or from dominance alone. Further structural properties of $\mathcal{M}_{enum}^{semi}$ have to be employed.
- 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 $\xi \xrightarrow{\mu/\xi} \mu$ and $\mathcal{M} \ni \xi$ for which $\xi \not\xrightarrow{\mu/\xi} \mu$.

Conclusions

- We discussed general mixture distributions and the important universality property – multiplicative dominance.
- We defined 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 reduced the analysis of these 49 cases to four basic cases.
- Domination (essentially by M) is known to be true for two cases. The remaining two (new) cases do not allow for domination.
- We improved the result on posterior convergence in ratio $\xi/\mu \to 1$ by providing the speed of convergence.
- We investigated whether convergence for all Martin-Löf random sequences could hold.

[http://www.idsia.ch/~marcus]