COMP4620/8620: Advanced Topics in AI
Foundations of Artificial Intelligence

Marcus Hutter
Australian National University
Canberra, ACT, 0200, Australia
http://www.hutter1.net/

ANU

4 ALGORITHMIC PROBABILITY & UNIVERSAL INDUCTION

- The Universal a Priori Probability M
- Universal Sequence Prediction
- Universal Inductive Inference
- Martin-Löf Randomness
- Discussion

Algorithmic Probability & Universal Induction: Abstract

Solomonoff completed the Bayesian framework by providing a rigorous, unique, formal, and universal choice for the model class and the prior. I will discuss in breadth how and in which sense universal (non-i.i.d.) sequence prediction solves various (philosophical) problems of traditional Bayesian sequence prediction. I show that Solomonoff's model possesses many desirable properties: Strong total and weak instantaneous bounds , and in contrast to most classical continuous prior densities has no zero p(oste)rior problem, i.e. can confirm universal hypotheses, is reparametrization and regrouping invariant, and avoids the old-evidence and updating problem. It even performs well (actually better) in non-computable environments.

Problem Setup

- Since our primary purpose for doing induction is to forecast (time-series), we will concentrate on sequence prediction tasks.
- Classification is a special case of sequence prediction.
  (With some tricks the other direction is also true)
- This Course focusses on maximizing profit (minimizing loss).
  We're not (primarily) interested in finding a (true/predictive/causal) model.
- Separating noise from data is not necessary in this setting!

Philosophy & Notation

Occam's razor: take simplest hypothesis consistent with data.

Epicurus' principle of multiple explanations: Keep all theories consistent with the data.

$\Downarrow$ $\Downarrow$

We now combine both principles:

Take all consistent explanations into account,
but weight the simpler ones higher.

Formalization with Turing machines and Kolmogorov complexity

Additional notation: We denote binary strings of length $\ell(x) = n$ by $x = x_{1:n} = x_1 x_2 ... x_n$ with $x_t \in \mathbb{B}$ and further abbreviate $x_{<n} := x_1 ... x_{n-1}$.

4.1 The Universal a Priori Probability M: Contents

- The Universal a Priori Probability M
- Relations between Complexities
- (Semi)Measures
- Sample Space / σ-Algebra / Cylinder Sets
- M is a SemiMeasure
- Properties of Enumerable Semimeasures
- Fundamental Universality Property of M

The Universal a Priori Probability M

Solomonoff defined the universal probability distribution $M(x)$ as the probability that the output of a universal monotone Turing machine starts with x when provided with fair coin flips on the input tape.

Definition 4.1 (Solomonoff distribution) Formally,

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

The sum is over minimal programs p for which U outputs a string starting with x (see Definition 2.6).

Since the shortest programs p dominate the sum, $M(x)$ is roughly $2^{-Km(x)}$. More precisely ...

Relations between Complexities

Theorem 4.2 (Relations between Complexities)

$KM := -\log M$, Km, and K are ordered in the following way:

$$0 \;\leq\; K(x|\ell(x)) \;\stackrel{+}{<}\; KM(x) \;\leq\; Km(x) \;\leq\; K(x) \;\stackrel{+}{<}\; \ell(x) + 2\log\ell(x)$$

Proof sketch:

The second inequality follows from the fact that,
given n and Kraft's inequality $\sum_{x\in\mathcal{X}^n} M(x) \leq 1$,
there exists for $x \in \mathcal{X}^n$ a Shannon-Fano code of length $-\log M(x)$,
which is effective since M is enumerable.

Now use Theorem 2.17 conditioned to n.

The other inequalities are obvious from the definitions. ■

(Semi)Measures

Before we can discuss the stochastic properties of M we need the concept of (semi)measures for strings.

Definition 4.3 ((Semi)measures) $\rho(x)$ denotes the probability that a binary sequence starts with string x. We call $\rho \geq 0$ a semimeasure if $\rho(\epsilon) \leq 1$ and $\rho(x) \geq \rho(x0) + \rho(x1)$, and a probability measure if equality holds.

The reason for calling ρ with the above property a probability measure is that it satisfies Kolmogorov's Axioms Definition 3.1 of probability in the following sense ...

Sample Space / Events / Cylinder Sets

- The The sample space is $\Omega = \mathbb{B}^\infty$ with elements $\omega = \omega_1\omega_2\omega_3... \in \mathbb{B}^\infty$ being infinite binary sequences.

- 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 ρ is uniquely defined by giving its values $\rho(\Gamma_{x_{1:n}})$ on the cylinder sets, which we abbreviate by $\rho(x_{1:n})$.

- We will also call ρ a measure, or even more loosely a probability distribution.

M is a SemiMeasure

- The reason for extending the definition to semimeasures is that M itself is unfortunately **not** a probability measure.
- 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.
- Since $M(\epsilon) = 1$, M **is** at least a semimeasure.

Properties of (Semi)Measure ρ

- Properties of ρ: $\sum_{x_{1:n}\in\mathcal{X}^n} \rho(x_{1:n}) \overset{(<)}{=} 1,$

$$\begin{aligned}\rho(x_t|x_{<t}) &:= \rho(x_{1:t})/\rho(x_{<t}),\\ \rho(x_1...x_n) &= \rho(x_1)\cdot\rho(x_2|x_1)\cdot...\cdot\rho(x_n|x_1...x_{n-1}).\end{aligned}$$

- One can show that ρ is an enumerable semimeasure

$$\Longleftrightarrow \exists \text{ mTM } T \;:\; \rho(x) = \sum_{p\,:\,T(p)=x*} 2^{-\ell(p)} \quad \text{and} \quad \ell(T) \overset{+}{=} K(\rho)$$

- Intuition: Fair coin flips are sufficient to create any probability distribution.

- Definition: $K(\rho) :=$ length of shortest self-delimiting code of a Turing machine computing function ρ in the sense of Def. 2.21.

Fundamental Universality Property of M

Theorem 4.4 (Universality of M)
M is a universal semimeasure in the sense that
$M(x) \stackrel{\times}{>} 2^{-K(\rho)} \cdot \rho(x)$ for all enumerable semimeasures ρ.
M is enumerable, but not estimable.

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

Proof sketch:

$$M(x) = \sum_{p\,:\,U(p)=x*} 2^{-\ell(p)} \geq \sum_{q\,:\,U(Tq)=x*} 2^{-\ell(Tq)} = 2^{-\ell(T)} \sum_{q\,:\,T(q)=x*} 2^{-\ell(q)} \stackrel{\times}{=} 2^{-K(\rho)}\rho(x)$$

4.2 Universal Sequence Prediction: Contents

Solomonoff, Occam, Epicurus

- In which sense does M incorporate Occam's razor and Epicurus' principle of multiple explanations?
- From $M(x) \approx 2^{-K(x)}$ we see that M assigns high probability to simple strings (Occam).
- More useful is to think of x as being the observed history.
- We see from Definition 4.1 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).

Prediction

How does all this affect prediction?

If $M(x)$ correctly 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 $K(x,y) \stackrel{+}{=} K(y|x,K(x)) + K(x)$ (Theorem 2.15), and assuming $K(x,y) \approx K(xy)$, and approximating $K(y|x,K(x)) \approx K(y|x)$, $M(x) \approx 2^{-K(x)}$, 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).

Simple Deterministic Bound

Sequence prediction algorithms try to predict the continuation $x_t \in \mathbb{B}$ of a given sequence $x_1...x_{t-1}$. Simple deterministic bound:

$$\sum_{t=1}^{\infty} |1-M(x_t|x_{<t})| \overset{a}{\leq} -\sum_{t=1}^{\infty} \ln M(x_t|x_{<t}) \overset{b}{=} -\ln M(x_{1:\infty}) \overset{c}{\leq} Km(x_{1:\infty}) \ln 2$$

(a) use $|1-a| \leq -\ln a$ for $0 \leq a \leq 1$.

(b) exchange sum with logarithm and eliminate product by chain rule.

(c) used Theorem 4.2.

If $x_{1:\infty}$ is a computable sequence, then $Km(x_{1:\infty})$ is finite,
which implies $M(x_t|x_{<t}) \to 1 \quad (\sum_{t=1}^{\infty} |1-a_t| < \infty \Rightarrow a_t \to 1)$.

$\Rightarrow$ if the environment is a computable sequence (digits of π or e or ...), 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.

Solomonoff's Major Result

Assume sequence $x_{1:\infty}$ is sampled from the unknown distribution μ, i.e. the true objective probability of $x_{1:n}$ is $\mu(x_{1:n})$.

The probability of x_t given $x_{<t}$ hence is $\mu(x_t|x_{<t}) = \mu(x_{1:t})/\mu(x_{<t})$.

Solomonoff's central result [Hut05] is that M converges to μ.

More precisely, he showed that

Theorem 4.5 (Predictive Convergence of M)

$$\sum_{t=1}^{\infty}\sum_{x_{<t}\in\mathbb{B}^{t-1}}\mu(x_{<t})\Big(M(0|x_{<t})-\mu(0|x_{<t})\Big)^2 \stackrel{+}{<} \tfrac{1}{2}\ln 2\cdot K(\mu) < \infty$$

Implications of Solomonoff's Result

- 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.
- Convergence is rapid: The expected number of times t in which $|M(0|x_{<t}) - \mu(0|x_{<t})| > \varepsilon$ 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 $c \stackrel{+}{=} \ln 2 \cdot K(\mu)$.
- No statement is possible for which t these deviations occur.
- This holds for **any** computable probability distribution μ.
- How does M know to which μ?
  The set of μ-random sequences differ for different μ.
- Intuition: Past data $x_{<t}$ are exploited to get a (with $t \to \infty$) improving estimate $M(x_t|x_{<t})$ of $\mu(x_t|x_{<t})$.
- Fazit: M is universal predictor. The only assumption made is that data are generated from a computable distribution.

Entropy Inequality

Proof of Solomonoff's bound: We need (proof as exercise)

Lemma 4.6 (Entropy Inequality)
$2(z-y)^2 \leq y\ln\frac{y}{z} + (1-y)\ln\frac{1-y}{1-z}$ for $0 < z < 1$ and $0 \leq y \leq 1$.

$$\leq y\ln\tfrac{y}{z} + (1-y)\ln\tfrac{1-y}{c-z} \quad \text{for} \quad 0 < z < c \leq 1 \quad \text{and} \quad 0 \leq y \leq 1.$$

The latter inequality holds, since the r.h.s. is decreasing in c. Inserting

$$0 \leq y := \mu(0|x_{<t}) = 1 - \mu(1|x_{<t}) \leq 1 \quad \text{and}$$

$$0 < z := M(0|x_{<t}) < c := M(0|x_{<t}) + M(1|x_{<t}) < 1 \quad \text{we get}$$

$$2(M(0|x_{<t}) - \mu(0|x_{<t}))^2 \;\leq\; \sum_{x_t\in\mathbb{B}} \mu(x_t|x_{<t})\ln\frac{\mu(x_t|x_{<t})}{M(x_t|x_{<t})} =: d_t(x_{<t})$$

The r.h.s. is the relative entropy between μ and M.

Proof of the Entropy Bound

$$D_n(\mu||M) \equiv \sum_{t=1}^{n}\sum_{x_{<t}} \mu(x_{<t})\cdot d_t(x_{<t}) \overset{(a)}{=} \sum_{t=1}^{n}\sum_{x_{1:t}} \mu(x_{1:t}) \ln \frac{\mu(x_t|x_{<t})}{M(x_t|x_{<t})} =$$

$$\overset{(b)}{=} \sum_{x_{1:n}} \mu(x_{1:n}) \ln \prod_{t=1}^{n} \frac{\mu(x_t|x_{<t})}{M(x_t|x_{<t})} \overset{(c)}{=} \sum_{x_{1:n}} \mu(x_{1:n}) \ln \frac{\mu(x_{1:n})}{M(x_{1:n})} \overset{(d)}{\overset{+}{<}} K(\mu) \ln 2$$

(a) Insert def. of d_t and use product rule $\mu(x_{<t})\cdot\mu(x_t|x_{<t})=\mu(x_{1:t})$.

(b) $\sum_{x_{1:t}} \mu(x_{1:t}) = \sum_{x_{1:n}} \mu(x_{1:n})$ and argument of log is independent of $x_{t+1:n}$. The t sum can now be exchanged with the $x_{1:n}$ sum and transforms to a product inside the logarithm.

(c) Use chain rule again for μ and M.

(d) Use dominance $M(x) \overset{\times}{>} 2^{-K(\mu)}\mu(x)$.

Inserting d_t into D_n yields Solomonoff's Theorem 4.5. ■

4.3 Universal Inductive Inference: Contents

- Bayesian Sequence Prediction and Confirmation
- The Universal Prior
- The Problem of Zero Prior
- Reparametrization and Regrouping Invariance
- Universal Choice of Class $\mathcal{M}$
- The Problem of Old Evidence / New Theories
- Universal is Better than Continuous $\mathcal{M}$
- More Bounds / Critique / Problems

Bayesian Sequence Prediction and Confirmation

- Assumption: Sequence $\omega \in \mathcal{X}^\infty$ is sampled from the "true" probability measure μ, i.e. $\mu(x) := \mathbf{P}[x|\mu]$ is the μ-probability that ω starts with $x \in \mathcal{X}^n$.
- Model class: We assume that μ is unknown but known to belong to a countable class of environments=models=measures $\mathcal{M} = \{\nu_1, \nu_2, ...\}$. [no i.i.d./ergodic/stationary assumption]
- Hypothesis class: $\{H_\nu : \nu \in \mathcal{M}\}$ forms a mutually exclusive and complete class of hypotheses.
- Prior: $w_\nu := \mathbf{P}[H_\nu]$ is our prior belief in H_ν

$\Rightarrow$ Evidence: $\xi(x) := \mathbf{P}[x] = \sum_{\nu\in\mathcal{M}} \mathbf{P}[x|H_\nu]\mathbf{P}[H_\nu] = \sum_\nu w_\nu \nu(x)$ must be our (prior) belief in x.

$\Rightarrow$ Posterior: $w_\nu(x) := \mathbf{P}[H_\nu|x] = \frac{\mathbf{P}[x|H_\nu]\mathbf{P}[H_\nu]}{\mathbf{P}[x]}$ is our posterior belief in ν (Bayes' rule).

The Universal Prior

- Quantify the complexity of an environment ν or hypothesis H_ν by its Kolmogorov complexity $K(\nu)$.
- Universal prior: $w_\nu = \boxed{w_\nu^U := 2^{-K(\nu)}}$ is a decreasing function in the model's complexity, and sums to (less than) one.

$\Rightarrow$ $D_n(\mu||\xi) \leq K(\mu)\ln 2$, i.e. the number of ε-deviations of ξ from μ is proportional to the complexity of the environment.

- No other semi-computable prior leads to better prediction (bounds).
- For continuous $\mathcal{M}$, we can assign a (proper) universal prior (not density) $w_\theta^U = 2^{-K(\theta)} > 0$ for computable θ, and 0 for uncomp. θ.
- This effectively reduces $\mathcal{M}$ to a discrete class $\{\nu_\theta \in \mathcal{M} : w_\theta^U > 0\}$ which is typically dense in $\mathcal{M}$.
- This prior has many advantages over the classical prior (densities).

The Problem of Zero Prior

= the problem of confirmation of universal hypotheses

Problem: If the prior is zero, then the posterior is necessarily also zero.

Example: Consider the hypothesis $H = H_1$ that all balls in some urn or all ravens are black (=1) or that the sun rises every day.

Starting with a prior density as $w(\theta) = 1$ implies that prior $\mathbf{P}[H_\theta] = 0$ for all θ, hence posterior $P[H_\theta|1..1] = 0$, hence H never gets confirmed.

3 non-solutions: define $H = \{\omega = 1^\infty\}$ | use finite population | abandon strict/logical/all-quantified/universal hypotheses in favor of soft hyp.

Solution: Assign non-zero prior to $\theta = 1 \quad \Rightarrow \quad \mathbf{P}[H|1^n] \to 1$.

Generalization: Assign non-zero prior to all "special" θ, like $\frac{1}{2}$ and $\frac{1}{6}$, which may naturally appear in a hypothesis, like "is the coin or die fair".

Universal solution: Assign non-zero prior to all comp. θ, e.g. $w_\theta^U = 2^{-K(\theta)}$

Reparametrization Invariance

- New parametrization e.g. $\psi = \sqrt{\theta}$, then the ψ-density $\tilde{w}(\psi) = 2\sqrt{\theta}\, w(\theta)$ is no longer uniform if $w(\theta) = 1$ is uniform
  $\Rightarrow$ indifference principle is not reparametrization invariant (RIP).

- Jeffrey's and Bernardo's principle satisfy RIP w.r.t. differentiable bijective transformations $\psi = f^{-1}(\theta)$.

- The universal prior $w_\theta^U = 2^{-K(\theta)}$ also satisfies RIP w.r.t. simple computable f. (within a multiplicative constant)

Regrouping Invariance

- Non-bijective transformations:
  E.g. grouping ball colors into categories black/non-black.
- No classical principle is regrouping invariant.
- Regrouping invariance is regarded as a very important and desirable property. [Walley's (1996) solution: sets of priors]
- The universal prior $w_\theta^U = 2^{-K(\theta)}$ is invariant under regrouping, and more generally under all simple [computable with complexity $O(1)$] even non-bijective transformations. (within a multiplicative constant)
- Note: Reparametrization and regrouping invariance hold for arbitrary classes and are not limited to the i.i.d. case.

Universal Choice of Class $\mathcal{M}$

- The larger $\mathcal{M}$ the less restrictive is the assumption $\mu \in \mathcal{M}$.
- The class $\mathcal{M}_U$ of all (semi)computable (semi)measures, although only countable, is pretty large, since it includes all valid physics theories. Further, ξ_U is itself semi-computable [ZL70].
- Solomonoff's universal prior $M(x) :=$ probability that the output of a universal TM U with random input starts with x.
- Formally: $\boxed{M(x) := \sum_{p\,:\,U(p)=x*} 2^{-\ell(p)}}$ where the sum is over all (minimal) programs p for which U outputs a string starting with x.
- M may be regarded as a $2^{-\ell(p)}$-weighted mixture over all deterministic environments ν_p. ($\nu_p(x) = 1$ if $U(p) = x*$ and 0 else)
- $M(x)$ coincides with $\xi_U(x)$ within an irrelevant multiplicative constant.

The Problem of Old Evidence / New Theories

- What if some evidence $E \widehat{=} x$ (e.g. Mercury's perihelion advance) is known well-before the correct hypothesis/theory/model $H \widehat{=} \mu$ (Einstein's general relativity theory) is found?
- How shall H be added to the Bayesian machinery a posteriori?
- What should the "prior" of H be?
- Should it be the belief in H in a hypothetical counterfactual world in which E is not known?
- Can old evidence E confirm H?
- After all, H could simply be constructed/biased/fitted towards "explaining" E.

Solution of the Old-Evidence Problem

- The universal class $\mathcal{M}_U$ and universal prior w_ν^U formally solves this problem.
- The universal prior of H is $2^{-K(H)}$ independent of $\mathcal{M}$ and of whether E is known or not.
- Updating $\mathcal{M}$ is unproblematic, and even not necessary when starting with $\mathcal{M}_U$, since it includes all hypothesis (including yet unknown or unnamed ones) a priori.

Universal is Better than Continuous $\mathcal{M}$

- Although $\nu_\theta()$ and w_θ are incomp. for cont. classes $\mathcal{M}$ for most θ, $\xi()$ is typically computable. (exactly as for Laplace or numerically)

$$\Rightarrow \boxed{D_n(\mu||M) \stackrel{+}{<} D_n(\mu||\xi)+K(\xi)\ln 2 \text{ for all } \mu}$$

- That is, M is superior to all computable mixture predictors ξ based on any (continuous or discrete) model class $\mathcal{M}$ and weight $w(\theta)$, save an additive constant $K(\xi)\ln 2 = O(1)$, even if environment μ is not computable.

- While $D_n(\mu||\xi) \sim \frac{d}{2}\ln n$ for all $\mu \in \mathcal{M}$, $D_n(\mu||M) \leq K(\mu)\ln 2$ is even finite for computable μ.

Fazit: Solomonoff prediction works also in non-computable environments

Convergence and Bounds

- Total (loss) bounds: $\sum_{n=1}^{\infty}\mathbb{E}[h_n] \stackrel{+}{<} K(\mu)\ln 2$, where $h_t(\omega_{<t}) := \sum_{a\in\mathcal{X}}(\sqrt{\xi(a|\omega_{<t})}-\sqrt{\mu(a|\omega_{<t})})^2$.
- Instantaneous i.i.d. bounds: For i.i.d. $\mathcal{M}$ with continuous, discrete, and universal prior, respectively:

  $\mathbb{E}[h_n] \stackrel{\times}{<} \frac{1}{n}\ln w(\mu)^{-1}$ and $\mathbb{E}[h_n] \stackrel{\times}{<} \frac{1}{n}\ln w_\mu^{-1} = \frac{1}{n}K(\mu)\ln 2$.
- Bounds for computable environments: Rapidly $M(x_t|x_{<t})\to 1$ on every computable sequence $x_{1:\infty}$ (whichsoever, e.g. 1^∞ or the digits of π or e), i.e. M quickly recognizes the structure of the sequence.
- Weak instantaneous bounds: valid for all n and $x_{1:n}$ and $\bar{x}_n\neq x_n$: $2^{-K(n)} \stackrel{\times}{<} M(\bar{x}_n|x_{<n}) \stackrel{\times}{<} 2^{2Km(x_{1:n})-K(n)}$
- Magic instance numbers: e.g. $M(0|1^n) \stackrel{\times}{=} 2^{-K(n)} \to 0$, but spikes up for simple n. M is cautious at magic instance numbers n.
- Future bounds / errors to come: If our past observations $\omega_{1:n}$ contain a lot of information about μ, we make few errors in future: $\sum_{t=n+1}^{\infty}\mathbb{E}[h_t|\omega_{1:n}] \stackrel{+}{<} [K(\mu|\omega_{1:n})+K(n)]\ln 2$

More Stuff / Critique / Problems

- Prior knowledge y can be incorporated by using "subjective" prior $w^U_{\nu|y} = 2^{-K(\nu|y)}$ or by prefixing observation x by y.
- Additive/multiplicative constant fudges and U-dependence is often (but not always) harmless.
- Incomputability: K and M can serve as "gold standards" which practitioners should aim at, but have to be (crudely) approximated in practice (MDL [Ris89], MML [Wal05], LZW [LZ76], CTW [WST95], NCD [CV05]).

4.4 Martin-Löf Randomness: Contents

- When is a Sequence Random? If it is incompressible!
- Motivation: For a fair coin 00000000 is as likely as 01100101, but we "feel" that 00000000 is less random than 01100101.
- Martin-Löf randomness captures the important concept of randomness of **individual** sequences.
- Martin-Löf random sequences pass all effective randomness tests.

When is a Sequence Random?

a) Is 01100101001011011010011111011 generated by a fair coin flip?
b) Is 11111111111111111111111111111 generated by a fair coin flip?
c) Is 11001001000011111110110101010 generated by a fair coin flip?
d) Is 01010101010101010101010101010 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 $\Longrightarrow$ (a,c,d) are fair Bernoulli($\frac{1}{2}$) coins, (b) not.
- Problem: (d) is non-random, also (c) is binary expansion of π.
- Solution: Choose H larger, but how large? Overfitting? MDL?
- AIT Solution: A sequence is **random** *iff* it is **incompressible**.

Martin-Löf Random Sequences

Characterization equivalent to Martin-Löf's original definition:

Theorem 4.7 (Martin-Löf random sequences)
A sequence $x_{1:\infty}$ is μ-random (in the sense of Martin-Löf)
$\iff$ there is a constant c such that $M(x_{1:n}) \leq c \cdot \mu(x_{1:n})$ for all n.

Equivalent formulation for computable μ:

$$x_{1:\infty} \text{ is } \mu\text{.M.L.-random} \quad \iff \quad Km(x_{1:n}) \stackrel{+}{=} -\log\mu(x_{1:n}) \; \forall n, \quad (4.8)$$

Theorem 4.7 follows from (4.8) by exponentiation, "using $2^{-Km} \approx M$" and noting that $M \stackrel{\times}{>} \mu$ follows from universality of M.

Properties of ML-Random Sequences

- Special case of μ being a fair coin, i.e. $\mu(x_{1:n}) = 2^{-n}$, then $x_{1:\infty}$ is random $\iff Km(x_{1:n}) \stackrel{+}{=} n$, i.e. iff $x_{1:n}$ is incompressible.
- For general μ, $-\log\mu(x_{1:n})$ is the length of the Arithmetic code of $x_{1:n}$, hence $x_{1:\infty}$ is μ-random $\iff$ the Arithmetic code is optimal.
- One can show that a μ-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-random sequences has μ-measure 1.

4.5 Discussion: Contents

- Limitations of Other Approaches
- Summary
- Exercises
- Literature

Limitations of Other Approaches 1

- Popper's philosophy of science is seriously flawed:
  – falsificationism is too limited,
  – corroboration $\equiv$ confirmation or meaningless,
  – simple $\neq$ easy-to-refute.
- No free lunch myth relies on unrealistic uniform sampling.
  Universal sampling permits free lunch.
- Frequentism: definition circular,
  limited to i.i.d. data, reference class problem.
- Statistical Learning Theory: Predominantly considers i.i.d. data:
  Empirical Risk Minimization, PAC bounds, VC-dimension,
  Rademacher complexity, Cross-Validation.

Limitations of Other Approaches 2

- Subjective Bayes: No formal procedure/theory to get prior.
- Objective Bayes: Right in spirit, but limited to small classes unless community embraces information theory.
- MDL/MML: practical approximations of universal induction.
- Pluralism is globally inconsistent.
- Deductive Logic: Not strong enough to allow for induction.
- Non-monotonic reasoning, inductive logic, default reasoning do not properly take uncertainty into account.
- Carnap's confirmation theory: Only for exchangeable data. Cannot confirm universal hypotheses.
- Data paradigm: Data may be more important than algorithms for "simple" problems, but a "lookup-table" AGI will not work.
- Eliminative induction ignores uncertainty and information theory.

Summary

- Solomonoff's universal a priori probability $M(x)$

  $=$ Occam $+$ Epicurus $+$ Turing $+$ Bayes $+$ Kolmogorov

  $=$ output probability of a universal TM with random input

  $=$ enum. semimeasure that dominates all enum. semimeasures

  $\approx 2^{-\text{Kolmogorov complexity}(x)}$

- $M(x_t|x_{<t}) \to \mu(x_t|x_{<t})$ rapid w.p.1 $\forall$ computable μ.
- M solves/avoids/meliorates many if not all philosophical and statistical problems around induction.
- Fazit: M is universal predictor.
- Matin-Löf /Kolmogorov define randomness of individual sequences:

  A sequence is random *iff* it is incompressible.

Exercises

1. [C10] Show that Definition 4.1 of M and the one given above it are equivalent.

2. [C30] Prove that ρ is an enumerable semimeasure if and only if there exists a TM T with $\rho(x) = \sum_{p:T(p)=x*} 2^{-\ell(p)} \; \forall x$.

3. [C10] Prove the bounds of Theorem 4.2

4. [C15] Prove the entropy inequality Lemma 4.6.
   Hint: Differentiate w.r.t. z and consider $y < z$ and $y > z$ separately.

5. [C10] Prove the claim about (rapid) convergence after Theorem 4.5 (Hint: Markov-Inequality).

6. [C20] Prove the instantaneous bound $M(1|0^n) \stackrel{\times}{=} 2^{-K(n)}$.

Literature

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

[LV08] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, Berlin, 3rd edition, 2008.

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005. http://www.hutter1.net/ai/uaibook.htm

[Hut07] M. Hutter. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007. http://arxiv.org/abs/0709.1516

[RH11] S. Rathmanner and M. Hutter. A philosophical treatise of universal induction. *Entropy*, 16(6):1076–1136, 2011. http://dx.doi.org/10.3390/e13061076