Kolmogorov Complexity - Computability (cont'd)

Proof idea intuitively: define q as the

> shortest string than cannot be specified with fewer than twelve words.

But q has just been specified with eleven words!

This paradox implies that the property *can be specified with n words* is not decidable.

The property is analogical to $K(q) < n$ where $n = 12$ and 'specified with n words' means being output by a universal T.M. with a program of length n.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Kolmogorov Complexity - Enumerability

A function $f(x) : \mathbb{N} \to \mathbb{R}$ is **enumerable** if there is a Turing machine finitely computing a function $f(x, k)$ *(i.e., halts for each x,k written in the binary form on the input tape and writes $f(x, k)$ in the binary form on the output tape)* such that

- $\lim_{k\to\infty} f(x, k) = f(x)$
- $f(x, k) \leq f(x, k + 1)$

for $\forall k$. A function is **co-enumerable** under the same conditions except $\leq$ is replaced by $\geq$.

Theorem 1

K is co-enumerable.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Kolmogorov Complexity - Enumerability (cont'd)

Proof of Theorem 1: Use the following algorithm

1. $k := 1$, $K(x, k) := |q| + c$ (??). Start *all programs shorter than* $|q| + c$ in parallel.
2. $k := k + 1$. Make one step in all running programs.
   - If any program halts and has produced q, select the shortest among them and set $K(x, k)$ to its length; stop all programs of length $K(x, k)$ or greater;
   - otherwise $K(x, k) := K(x, k - 1)$

   Go to 2.

Some of the programs started in 1 will never halt so this procedure will neither, and we will never know how close $K(q, k)$ is to $K(q)$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Unknown Environment

With zero knowledge about the environment, consider assigning *higher probabilities to simpler* x_k, i.e., sequences with with lower $K(x_k)$.

So e.g.

$$P(1|11111111) > P(0|11111111)$$

because $K(111111111) < K(111111110)$.

Note: $P(x_k|x_{<k}) = c \cdot P(x_{\leq k})$ *where* $c = 1/P(x_{<k})$ *is a constant after observing* $x_{<k}$.

The problem is that $K(x_{\leq k})$ depends only on the single shortest program generating $x_{\leq k}$. Intuitively, the probability of $x_{\leq k}$ should be higher if there are multiple simple programs each generating $x_{\leq k}$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Universal Prior M

Solomonoff (1964) proposed the *universal prior* which is closely related to K but accounts for multiple generating programs:

$$M(x_{\leq k}) = \sum_{p:U(p)=x_{\leq k}*} 2^{-|p|} \tag{1}$$

where the sum is over all *minimal* programs for which the universal T.M. U outputs a string starting with $x_{\leq k}$, not necessarily halting.

(Minimal program = all the has been read from the input tape when all of $x_{\leq k}$ has appeared on the output tape. Ignores any 'useless' continuation of the input.)

So all minimal programs p generating $x_{\leq k}$ contribute to $x_{\leq k}$'s probability but short programs contribute exponentially more than long programs.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Universal Prior M: Properties

$M(x_{\leq k})$ is close to $2^{-K(x_{\leq k})}$ since the shortest program generating $x_{\leq k}$ contributes exponentially more to $M(x_{\leq k})$ than other programs.

M is enumerable (proof omitted).

A function P is called a **semi-measure** if it satisfies all probability axioms but in the countable-union axiom $\sum_{i=1}^{\infty} P(e_i) = P(\cup_{i=1}^{\infty} e_i)$, $=$ is replaced by $\leq$.

M is a semi-measure. Indeed e.g.

$$M(000) + M(001) < M(00)$$

since there are programs outputting 00 and halting or looping forever afterwards without writing 0 or 1.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Universal Prior M: Properties (cont'd)

Theorem 2

For any computable $x_{\leq k} \in X^*$,

$$\lim_{k \to \infty} M(x_k | x_{<k}) = 1 \tag{2}$$

This means that after "seeing" the beginning $x_{<k}$ of the sequence, M predicts the next element with probability approaching 1 with $k \to \infty$. So M "recognizes" the environment on the only condition that the latter produces a computable sequence, i.e., it is a Turing machine.

The condition above is **not** *strong: all physical theories of the world are computable, so any "reasonable" environment is a T.M. But recall the catch: M itself is not computable, only enumerable.*

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Proof of Theorem 2

Because $(1-a)^2 \leq -\frac{1}{2}\ln a$ (for $0 \leq a \leq 1$):

$$\sum_{k=1}^{\infty}(1-M(x_k|x_{<k}))^2 \leq -\frac{1}{2}\sum_{k=1}^{\infty}\ln M(x_k|x_{<k})$$

Swap the sum with the logarithm:

$$= -\frac{1}{2}\ln M(x_1)\cdot M(x_2|x_1)\cdot M(x_3|x_{<3})\cdot\ldots$$

Apply the chain-rule:

$$= -\frac{1}{2}\ln M(x_{1:\infty})$$

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Proof of Theorem 2 (cont'd)

Plug in the definition of M, then drop from the sum all p's computing $x_{1:\infty}$ except for the shortest one denoted $p_{\min}$

$$= -\frac{1}{2}\ln \sum_{p:U(p)=x_{1:\infty}} 2^{-|p|} \leq -\frac{1}{2}\ln 2^{-|p_{\min}|} \leq \frac{1}{2}\ln 2 \cdot |p_{\min}|$$

If $x_{1:\infty}$ is computable then clearly $|p_{\min}| < \infty$ and so

$$\sum_{k=1}^{\infty} (1 - M(x_k|x_{<k}))^2 < \infty \tag{3}$$

Finally, (2) must hold, otherwise 3 would not hold.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

The AIXI Agent

Let us now see how M can be used in the agent-environment setting involving not only observations but also rewards and actions.

Assume for simplicity that rewards and actions are binary $Y = R = \{0, 1\}$.
(No loss of generality; any finite description can be expressed as a binary string.)

The percept history $xr_{\leq k}$ is then a binary string so $M(xr_{\leq k})$ is defined by (1).

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

The AIXI Agent (cont'd)

To quantify the probability of percept history $xr_{\leq k}$ *given action history* $y_{\leq k}$, Hutter (2005) proposed to adapt (1) into

$$M(xr_{\leq m} \mid y_{\leq m}) = \sum_{p:U(p,y_{\leq m})=xr_{\leq k}*} 2^{-|p|} \tag{4}$$

where $m \in \mathbb{N}$ and the sum is over all programs for the universal T.M. which outputs $xr_{\leq m}$ (followed by any suffix) *given* $y_{\leq m}$ *on the input tape*.

The program p can be distinguished from its input $y_{\leq m}$ *on the input tape e.g. by the coding described here.*

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

The AIXI Agent (cont'd)

The agent which executes the sequence of actions

$$y_{\leq m} = \arg\max_{y_{\leq m}} \sum_{xr_{\leq m}} \left(M(xr_{\leq m} \mid y_{\leq m}) \sum_{k=1}^{m} r_k \right) \tag{5}$$

is the **AIXI agent**.

Theorem 3

The AIXI agent is Pareto-optimal with respect to maximizing utility (??), *i.e., there is no other agent that performs at least as well as AIXI in all environments while performing strictly better in at least one environment.*

Our presentation of AIXI is simplified. The resources here provide a proof for the theorem and also explain how the optimal action y_k is computed from the history $xr_{\leq k}, y_{\leq k}$ and how AIXI is defined for an infinite horizon ($m \to \infty$).

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

M as a Bayesian Mixture

Define $K(f)$ for a function $f : \mathbb{N} \to \mathbb{R}$ as the length of the shortest program computing *(a binary representation of)* $f(q)$ given *(a binary representation of)* input q.

Theorem 4

Let $\mathcal{M}_U$ be the set of all enumerable semi-measures. $M(q)$ (2) is equivalent to

$$\xi(q) = \sum_{p \in \mathcal{M}_U} p(q) \cdot 2^{-K(p)} \tag{6}$$

in the sense that

$$M(q) = \mathcal{O}(\xi(q)) \text{ and } \xi(q) = \mathcal{O}(M(q))$$

Note that $M \in \mathcal{M}_U$ because it is an enumerable semi-measure.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

M as a Bayesian Mixture (cont'd)

Online sequence prediction with ξ can be done with iterative updates: the model at k is

$$\xi_k(x_{k+1} \mid x_{\leq k}) = \sum_{p \in \mathcal{M}_U} p(x_{k+1} \mid x_{\leq k}) B_k(p) \tag{7}$$

the initial beliefs in models $p \in \mathcal{M}_U$ are

$$B_1(p) = 2^{-K(p)} \tag{8}$$

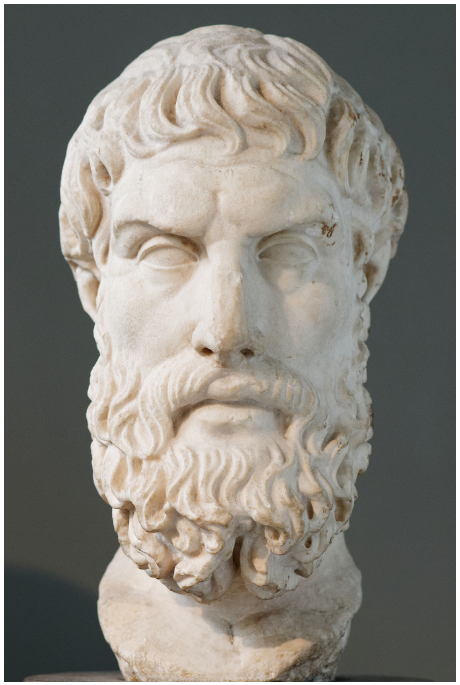
and they are updated (with the normalizer α) by

$$B_{k+1}(p) = \alpha p(x_{k+1} \mid x_{\leq k}) B_k(p) \tag{9}$$

(7) and (9) are analogical to (??) and (??), but here the model class $\mathcal{M}_U$ is the 'largest possible' and the initial beliefs are determined by (8).

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Principle of Multiple Explanations



Epicurus

Greek philosopher

cca 342-270 B.C.

Keep all explanations consistent with the observations.

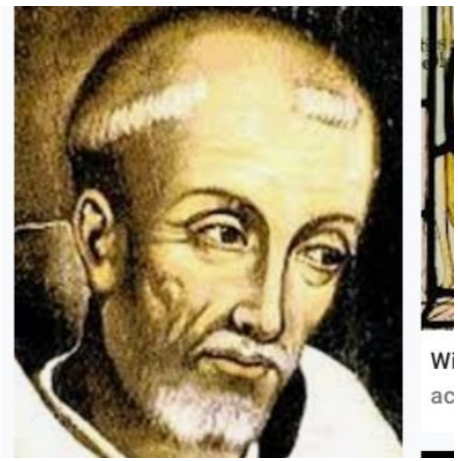
We have applied this principle in the version space algorithm.

But how to make a prediction using multiple explanations?

- In version space, we took the majority vote among all hypotheses. They all had the same voting weight.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Principle of Simplest Explanation



William of Ockham

English philosopher

cca 1290-1349 B.C.

Entities should not by multiplied beyond necessity

or

Keep the simplest explanation consistent with the observations.

Famously known as the Occam's razor

Kolmogorov complexity provides a mathematical way to determine 'simplest'.

But is it reasonable to discard all other explanations?

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Bayesian Inference

Thomas Bayes

English mathematician

cca 1701-1761 B.C.

ξ combines the two principles using Bayesian inference.

$$\xi(q) = \overbrace{\underbrace{\sum_{p \in \mathcal{M}_U}}_{\text{All models contribute. (Epicurus)}} p(q) \cdot \underbrace{2^{-K(p)}}_{\text{Simpler models get more weight. (Ockham)}}}^{\text{weighted mixture (Bayes)}}$$