Annals of Pure and Applied Logic 167 (2016) 897–900

Contents lists available at ScienceDirect

Annals of Pure and Applied Logic

www.elsevier.com/locate/apal

Occam bound on lowest complexity of elements ☆

Leonid A. Levin

Boston University, Computer Sci. dpt., 111 Cummington Mall, Boston, MA 02215, United States

ARTICLE INFO

Article history:
Available online 7 July 2016

MSC:
03B48
62C12
68Q30
68Q32

Keywords:
Occam
Prior
Extrapolation
Complexity
Randomness
Information

ABSTRACT

The combined universal probability $\mathbf{M}(D)$ of strings x in sets D is close to $\max_{x\in D}\mathbf{M}(\{x\})$: their $\sim$ logs differ by at most D's information $j{=}\mathbf{I}(D:\mathcal{H})$ about the halting sequence $\mathcal{H}$. Thus if all x have complexity $\mathbf{K}(x)\geq k$, D carries $\geq i$ bits of information on each x where $i{+}j\sim k$. Note, there are no ways (whether natural or artificial) to generate D with significant $\mathbf{I}(D:\mathcal{H})$.

1. Introduction

Many intellectual and computing tasks require guessing the hidden part of the environment from available observations. In different fields these tasks have various names, such as Inductive Inference, Extrapolation, Passive Learning, etc. The relevant part of the environment can be represented as an, often huge, string $x{\in}\{0,1\}^*$. The known observations restrict it to a set $D\ni x$.[1]

One popular approach to guessing, the "Occam Razor," tells to focus on the simplest members of D. (In words, attributed to A. Einstein, "A conjecture should be made as simple as it can be, but no simpler.") Its implementations vary: if two objects are close in simplicity, there may be legitimate disagreements on which is slightly simpler. This ambiguity is reflected in formalization of "simplicity" via the Kolmogorov

☆ This research was supported in part by NSF grant CCF-1049505.
URL: http://www.cs.bu.edu/fac/Lnd.

[1] D is typically enormous, and a much more concise theory can often represent the relevant part of what is known about x. Yet, such *ad hoc* approaches are secondary: raw observations are anyway their ultimate source.

http://dx.doi.org/10.1016/j.apal.2014.01.006
0168-0072/© 2016 Published by Elsevier B.V.

Complexity function $\mathbf{K}(x)$ – the length of the shortest prefix program[2] generating x: $\mathbf{K}$ is defined only up to an additive constant depending on the programming language. This constant is small compared to the usually huge whole bit-length of x. More mysterious is the justification of this Occam Razor principle.

A more revealing philosophy is based on the idea of "Prior". It assumes the guessing of $x \in D$ is done by restricting to D an *a priori* probability distribution on $\{0,1\}^*$. Again, subjective differences are reflected in ignoring moderate factors: say in asymptotic terms, priors different by $\theta(1)$ factors are treated as equivalent. The less we know about x (before observations restricting x to D) the more "spread" is the prior, i.e. the smaller would be the variety of sets that can be ignored due to their negligible probability. This means that distributions truly prior to any knowledge, would be the largest up to $\theta(1)$ factors. Among enumerable (i.e. generatable as outputs of randomized algorithms) distributions, such largest prior does in fact exist and is $\mathbf{M}(\{x\}) = 2^{-\mathbf{K}(x)}$.

These ideas developed in [9] and many subsequent papers do remove some mystery from the Occam Razor principle. Yet, they immediately yield a reservation: the simplest objects have **each** the highest universal probability, but it may still be negligible compared to the **combined** probability of complicated objects in D. This suggests that the general inference situation might be much more obscure than the widely believed Occam Razor principle describes it.

The present paper shows this could not happen, except as a purely mathematical construction. Any such D has high information $\mathbf{I}(D:\mathcal{H})$ about Halting Problem $\mathcal{H}$ ("Turing's Password" :-). So, they are "exotic": there are no ways to generate such D; see this informational version of Church–Turing Thesis discussed at the end of [5].

Consider finite sets D containing only strings of high ($\gtrsim k$) complexity. One way to find such D is to generate at random a small number of strings $x \in \{0,1\}^k$. With a little luck, all x would have high complexity, but D would contain virtually all information about each of them.

Another (less realistic :-) method is to gain access to the halting problem sequence $\mathcal{H}$ and use it to select for D all strings x of complexity $\sim k$ from among all k-bit strings. Then D contains little information about most of its x but much information about $\mathcal{H}$!

Yet another way is to combine both methods. Let v_h be the set of all strings vs with $\mathbf{K}(vs) \sim \|vs\| {=} \|v\| {+} h$. Then $\mathbf{K}(x) \sim i + h$, $\mathbf{I}(D:x) \sim i$, and $\mathbf{I}(D:\mathcal{H}) \sim h$ for most i-bit v and $x {\in} D = v_h$. We will see no D can be better: they all contain strings of complexity $\lesssim \min_{x\in D} \mathbf{I}(D:x) {+} \mathbf{I}(D:\mathcal{H})$.

The result is a follow-up to Theorem 2 in [11]. [10] provides in Appendix I more history of the concepts used here; [4,9,7] give more material on Algorithmic Information Theory. This work's central idea is due to S. Epstein, appearing in [2]. He is a co-author of an earlier preprint [3] of the results below and a sole author of their many extensions in [1].

2. Conventions and Kolmogorov complexity tools

$\|x\| \stackrel{\text{def}}{=} n$ for $x{\in}\{0,1\}^n$; for $a{\in}\Re^+$, $\|a\| \stackrel{\text{def}}{=} \lceil\, |\log a|\, \rceil$. $\mathbf{S} \stackrel{\text{def}}{=} \{0,1\}^*$. $p0^- {=} p1^- \stackrel{\text{def}}{=} p$; $\emptyset^-$ is undefined. $[A] \stackrel{\text{def}}{=} 1$ if statement A holds, else $[A] \stackrel{\text{def}}{=} 0$. ${\prec} f$, ${\succ} f$, ${\asymp} f$, and ${\lesssim} f$, ${\gtrsim} f$, ${\sim} f$ denote ${<} f {+} O(1)$, ${>} f {-} O(1)$, ${=} f {\pm} O(1)$, and ${<} f {+} O(\|f{+}1\|)$, ${>} f {-} O(\|f{+}1\|)$, ${=} f {\pm} O(\|f{+}1\|)$, respectively. $Q(G)$ is the probability of a set G or ***mean*** $\sum_x Q(\{x\}) G(x)$ of a function G by a distribution Q.

[2] This analysis ignores issues of finding short programs efficiently. Limited-space versions of absolute complexity results are usually straightforward. Time-limited versions often are not, due to difficulties of inverting one-way functions. However, the inversion problems have time-optimal algorithms. See such discussions in [6].