Philosophical Issues in Kolmogorov Complexity

Ming Li*
University of Waterloo

Paul M.B. Vitányi†
CWI and University of Amsterdam

1 Introduction

Five years have passed since we wrote our first survey on Kolmogorov complexity and its applications [10]. This essay is not meant to be an exhaustive survey of the subject, not even of the recent results; that is done thoroughly in our forthcoming book [13] which will appear very soon. Here, we would like to convey to our reader some appealing philosophical ideas by just picking up some pretty shells deposited on the shore by the sea of applications of Kolmogorov complexity. We hope these ideas will be useful or, at least, enjoyable to our reader.

We give preference to ideas and applications that were not (well) covered by our previous articles [10, 11], either due to our ignorance at the time or because the results are new. We also prefer those results that have deeper philosophical or methodological implications. During our narrative, we often venture into strange lands where we are only amateurs or even total strangers. Thus our views might not be completely conventional, but we do hope they are novel and interesting.

Due to space limitation, we refer the reader to [11, 13] for definitions and basic facts of Kolmogorov complexity. For the purpose of reading this article at a conceptual level, it is sufficient to know that Kolmogorov complexity of a finite string x is simply the length of the shortest program, say in FORTRAN[1] encoded in binary bits, which prints x without any input. $C(x)$ is the Kolmogorov complexity of x; $K(x)$ is the prefix Kolmogorov complexity of x where the program for x must be self-delimiting.

2 Should we prefer elementary proofs?

Probabilistic or information–theoretic style proofs have enjoyed major successes in combinatorics and computer science. Our thinking about proofs in computer science parallels the following comments of Kolmogorov [8] about information theory:

> The real substance of the entropy formula [based on probabilistic assumptions about independent random variables] ... holds under incomparably weaker

*Supported by NSERC operating grant OGP-046506.

Computer Science Department, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada. E-mail: mli@math.uwaterloo.ca

†Partially supported by NSERC International Scientific Exchange Award ISE0125663.

CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. E-mail: paulv@cwi.nl

[1] Or in Turing machine codes.

> and purely combinatorial assumptions... Information theory must precede probability theory, and not be based on it. By the very essence of this discipline, the foundations of information theory must have a finite combinatorial character.

From a practical viewpoint the real issue is whether elementary arguments must always be more tedious. We demonstrate through one example that elementary proofs (using Kolmogorov complexity) are not only more intuitive, but also easier. Kolmogorov complexity based arguments, although nonconstructive, are essentially combinatorial in nature without probabilistic assumptions. We use $d(S)$ to denote the number of elements in set S.

A family $\mathcal{D} = \{D_1, D_2, ..., D_j\}$ of subsets of $N = \{1, 2, \ldots, n\}$ is called a *distinguishing family* for N if for any two distinct subsets M and M' of N there exists an i $(1 \leq i \leq j)$ such that $d(D_i \cap M)$ is different from $d(D_i \cap M')$. Let $f(n)$ denote the minimum of $d(\mathcal{D})$ over all distinguishing families for N. The *coin-weighing problem* is to determine $f(n)$. It is known that

$$f(n) \leq (2n/\log n)[1 + O(\log\log n/\log n)]. \tag{1}$$

Equation 1 was independently established by B. Lindström in 1965 and D.G. Cantor and W.H. Mills in 1966. P. Erdös and A. Rényi [5], L. Moser [15], and N. Pippenger [16] have established the following Theorem 1 using various probabilistic and information theory methods (second moment method).

Fix an encoding of the 2^n subsets of N such that each subset is encoded by a binary string of length n. Simplifying notation, we write M as the encoding of M.

Theorem 1 $f(n) \geq (2n/\log n)[1 + O(\log\log n/\log n)]$.

PROOF. Choose M such that $C(M|D_2, ..., D_j) \geq n$. Let $d_i = d(D_i)$ and $m_i = d(D_i \cap M)$. By elementary estimates [13], m_i is within range $d_i/2 \pm O(\sqrt{d_i \log d_i})$. Thus, for $1 \leq i \leq j$, m_i can be described using its discrepancy with $d_i/2$, hence

$$C(m_i|D_i) \leq \frac{1}{2}\log d_i + O(\log\log d_i) \leq \frac{1}{2}\log n + O(\log\log n).$$

Since $\mathcal{D}$ is a distinguishing family, given $\mathcal{D}$, the values $m_1, ..., m_j$ determine M:

$$n \leq C(M|D_1, ..., D_j) \leq C(m_1, ..., m_j|D_1, ...D_j) \leq \Sigma_{i=1}^{j}(\frac{1}{2}\log n + O(\log\log n)).$$

This implies the theorem. □

3 The Grue Emerald Paradox

For about two thousand years philosophers have worried about the problem of inductive reasoning. On the one hand, it seems common sense to assume that people learn in the sense that they generalize from observations by learning a 'Law' that governs not only the past observations, but will also apply to the observations in the future. In this sense induction should 'add knowledge'.

Yet how is it possible to acquire knowledge which is not yet present? If we have a system to deduce a general law from observations, then this law is only part of the knowledge contained in this system and the observations. Then, the law does not represent knowledge over and above what was already present, but it represents in fact only a part of that knowledge.

In [7], N. Goodman described the *grue emerald paradox.* Let h be the hypothesis that all emeralds are green. Let k be the hypothesis that all emeralds examined before the the year of 2000 are green and all emeralds examined after 2000 will be blue. Goodman called this color 'grue'. Then both hypotheses are totally confirmed by the experiments so far. How do we develop some tools, or philosophy, to prefer h than k? People have been resorting to farfetched arguments, for example, to prefer time-independent hypothesis (h) than the time-dependent hypothesis (k).

Francis Bacon, in *Sylva Sylvarum* 337, 1627, formulates the power of induction as follows: "The eye of the understanding is like the eye of the sense; for as you may see great objects through small crannies or levels, so you may see great axioms of nature through small and contemptible instances."

Mathematics has come up with an induction principle which has an impeccable derivation, yet allows us to estimate the relative likelihood of different possible hypotheses—which is impossible with the commonly used Pearson-Neyman testing. Consider a discrete sample space Ω. Let $D, H_1, H_2, \ldots$ be a countable set of events (subsets) of Ω. $\mathbf{H} = \{H_1, H_2, \ldots\}$ is called *hypotheses space* . The hypotheses H_i are exhaustive (at least one is true). From the definition of conditional probability, that is, $P(A|B) = P(A \cap B)/P(B)$, it is easy to derive **Bayes' formula** (rewrite $P(A \cap B)$ in two different ways):

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)}. \tag{2}$$

If the hypotheses are mutually exclusive ($H_i \cap H_j = \emptyset$ for all i, j), then

$$P(D) = \sum_i P(D|H_i)P(H_i).$$

Despite the fact that Bayes' rule is just a rewriting of the definition of conditional probability and nothing more, it is its interpretation and applications that are most profound and caused much bitter controversy during the past two centuries. In Equation 2, the H_i's represent the possible alternative hypotheses concerning the phenomenon we wish to discover. The term D represents the empirically or otherwise known data concerning this phenomenon. The term $P(D)$, the probability of data D, may be considered as a normalizing factor so that $\sum_i P(H_i|D) = 1$. The term $P(H_i)$ is called the *a priori* probability or *initial* probability of hypothesis H_i, that is, it is the probability of H_i being true before we see any data. The term $P(H_i|D)$ is called *a posteriori* or *inferred* probability

The most interesting term is the prior probability $P(H_i)$. In the context of machine learning, $P(H_i)$ is often considered as the learner's *initial degree of belief* in hypothesis H_i. In essence Bayes' rule is a *mapping* from *a priori* probability $P(H_i)$ to *a posteriori* probability $P(H_i|D)$ determined by data D. In general, the problem is not so much that in the limit the inferred hypothesis would not concentrate on the true hypothesis, but that the inferred probability gives as much information as possible about the possible hypotheses from only a limited number of data. In fact, the continuous bitter debate between the Bayesian and non-Bayesian opinions centered on the prior probability. The controversy is caused by the fact that Bayesian theory does not say how to initially derive the prior probabilities for the hypotheses. Rather, Bayes' rule only tells how they are to be *updated.* In the real world problems, the prior proabilities may be unknown, uncomputable, or even conceivably non-existent. (What is the prior probability of use of a word in written English? There are many different sources of different social backgrounds living in different ages.) This problem would be solved if we can find a *single* probability distribution to use as the prior distribution in each different case, with approximately the same result as if we had used the real distribution. Surprisingly, this turns out to be possible up to some mild restrictions.

Consider theory formation in science as the process of obtaining a compact description of the past observations. The investigator observes increasingly larger initial segments of an infinite binary sequence X as the outcome of an infinite sequence of experiments on some aspect of nature. To describe the underlying regularity of X, the investigator tries to formulate a theory that governs X, consistent with past experiments. Candidate theories (hypotheses) are identified with computer programs that compute binary sequences starting with the observed initial segment.

First assume the existence of a prior probability distribution μ (actually a measure) over the continuous sample space $\Omega = \{0,1\}^\infty$. Denote by $\mu(x)$ the probability of a sequence starting with x. Given a previously observed data string S, the inference problem is to predict the next symbol in the output sequence, that is, extrapolating the sequence S. In terms of the variables in Equation 2, H_a is the hypothesis that the sequence under consideration continues with a. Data D_S consists of the fact that the the sequence starts with initial segment S. Thus, for $P(H_i)$ and $P(D)$ in Formula 2 we substitute $\mu(H_a)$ and $\mu(D_S)$, respectively, and obtain:

$$\mu(H_a|D_S) = \frac{\mu(D_S|H_a)\mu(H_a)}{\mu(D_S)}.$$

We must have $\mu(D_S|H_a) = 1$ for any a, hence,

$$\mu(H_a|D_S) = \frac{\mu(H_a)}{\mu(D_S)}. \tag{3}$$

Generally, we denote $\mu(H_a|D_S)$ by $\mu(a|S)$. In terms of inductive inference or machine learning, the final probability $\mu(a|S)$ is the probability of the next symbol being a, given the initial sequence S. Obviously we now only need the prior probability μ to evaluate $\mu(a|S)$.

The idea is to approximate the unknown proper prior probability μ. Without too much loss of generality we may as well assume that the measure μ is *enumerable.* That means, there is a Turing machine T which computes a total function $\phi(x,k)$ such that $\phi(x,k+1) \geq \phi(x,k)$ and $\lim_{k\to\infty}\phi(x,k) = \mu(x)$. If μ is recursive then it is also enumerable, but not necessarily the converse. It turns out that the class of all enumerable measures contains a *universal measure*, denoted by $\mathbf{M}$, such that for all μ in this class there exists a constant $c > 0$ such that $\mathbf{M}(x) \geq c\mu(x)$ for all x. We say that $\mathbf{M}$ *dominates* μ. We also call $\mathbf{M}$ the *a priori* probability, since it assigns maximal probability to all hypotheses in absence of any knowledge about them.

Now instead of using Formula 3, we estimate the conditional probability $\mu(y|x)$ that the next segment after x is y by the expression

$$\frac{\mathbf{M}(xy)}{\mathbf{M}(x)}. \tag{4}$$

Now let μ in Formula 3 be an arbitrary computable measure. This case includes all computable sequences. If the length of y is fixed, and the length of x grows to infinity, then it can be shown [18] that

$$\frac{\mathbf{M}(y)/\mathbf{M}(x)}{\mu(y)/\mu(x)} \rightarrow 1,$$

with μ-probability one. In other words, the conditional *a priori* probability is almost always asymptotically equal to the conditional probability. It has also shown by Solomonoff that the convergence is very fast and if we use Formula 4 instead of the real value Formula 3, then our inference is almost as good. We also know that

$$-\log \mathbf{M}(x) = K(x) + O(\log K(x)), \tag{5}$$

That means that $\mathbf{M}$ assigns high probability to simple objects and low probability to complex or random objects. We now come to the punch line: Bayes' rule using the universal prior distribution yields Occam's Razor principle. Namely, if several programs could generate $S0$ then the shortest one is used (for the prior probability), and further if $S0$ has a shorter program than $S1$ then $S0$ is preferred (that is, predict 0 with higher probability than predicting 1 after seeing S). Bayes' rule via the universal prior distribution also gives the so-called indifference principle in case $S0$ and $S1$ have roughly equal length shortest programs which 'explain' $S0$ and $S1$, respectively. The Goodman's grue emerald paradox disappears under this paradigm.

Scientists formulate their theories in two steps: firstly a scientist, based on scientific observations, formulate alternative hypotheses, and secondly a definite hypothesis is selected. The second step is the subject of inference in statistics. Statisticians have developed many different principles to do this, like Occam's Razor principle, the Maximum Likelihood principle, various ways of using Bayes' formula with different prior distributions. No single principle turned out to be satisfiable in all situations. Philosophically speaking, Solomonoff's approach presents an ideal way of solving induction problems. However, due to the non-computability of the universal prior function, such a theory cannot be directly used. Some approximation is needed in the real world applications.

Now we will closely follow Solomonoff's idea, but substitute a 'good' computable approximation to $\mathbf{M}(x)$. This results in Rissanen's **Minimum Description Length** principle [17]. Rissanen not only gives the principle, more importantly he also gives the detailed formulae on how to use this principle. This made it possible to use the MDL principle. The MDL principle can be intuitively stated as follows:

Minimum Description Length Principle. *The best theory to explain a set of data is the one which minimizes the sum of*

- *the length, in bits, of the description of the theory;*
- *the length, in bits, of data when encoded with the help of the theory.*

We now develop this MDL principle from Bayes' rule using the universal distribution $\mathbf{M}(x)$, assuming that P is enumerable. From the Bayes' Formula 2, we must choose the hypothesis H such that $P(H|D)$ is maximized. First we take the negative logarithm on both sides of Equation 2, we get

$$-\log P(H|D) = -\log P(D|H) - \log P(H) + \log P(D)$$

$\log P(D)$ is a constant and hence ignored. Maximizing the $P(H|D)$ over all possible H's is equivalent to *minimizing* $-\log P(H|D)$, or minimizing

$$-\log P(D|H) - \log P(H)$$

Now to get the minimum description length principle, we only need to explain above two terms in the sum properly. According to Solomonoff, when P is enumerable, then we approximate P by $\mathbf{M}$. The prior probability $P(H)$ is set to $\mathbf{M}(H) = 2^{-K(H) \pm O(\log K(H))}$, where $K(H)$ is the prefix-complexity of H. That is, $-\log P(H)$ is about the *length* of a minimum *prefix code*, or program, of hypothesis H.

A similar argument applies to term $-\log P(D|H)$. That is, $2^{-K(D|H) \pm O(\log K(D|H))}$ is a reasonable approximation of $P(D|H)$. The term $-\log P(D|H)$, also known as the *self-information* in information theory and the negative log likelihood in statistics, can now be regarded as the number of bits it takes to redescribe or encode D with an ideal code relative to H. In different applications, the hypothesis H can mean many different things, such as decision trees, finite automata, Boolean formulae, or a polynomial. In general statistical applications, one assumes that H is some model $H(\theta)$ with a set of parameters $\theta = \{\theta_1, \ldots, \theta_k\}$ of precision c, where the number k may vary and influence the descriptional complexity of $H(\theta)$. In such case, we minimize

$$-\log P(D|\theta) - \log P(\theta).$$

Let's consider one example. For each fixed k, $k = 0, \ldots, n-1$, let f_k be the best polynomial of degree k, fitted on points (x_i, y_i) $(1 \leq i \leq n)$, which minimizes the error

$$error(f_k) = \sum_{i=1}^{n} (f_k(x_i) - y_i)^2.$$

Assume each coefficient takes c bits. So f_k is encoded in ck bits. Let us assume the commonly used Gaussian distribution of the error on y_i's. Thus, given that f_k is the true polynomial,

$$\Pr(y_1, \ldots, y_n | f_k) = \Pi_i \exp(-O((f_k(x_i) - y_i)^2)).$$

The negative logarithm of above is $c' \cdot error(f_k)$ for some computable c'. The MDL principle tells us to choose f_k, $k \in \{0, \ldots, n-1\}$, which minimizes $ck + c' \cdot error(f_k)$.

4 Valiant learning under computable distributions?

Valiant's model [20] provides an excellent framework for studying learnability. Subsequent investigations show many problems intractable (NP-complete) under the original model. Can we adapt the it to obtain a model where more concepts are polynomial time learnable? The philosophy here is that maybe humans just learn a concept under *some restricted class of distributions*, like computable ones (those in our textbooks). Kolmogorov complexity and the Solomonoff-Levin universal distribution allows us to systematically develop a theory of Valiant-style learning under all (semi) computable distributions.

All distributions we have a name for: the uniform distribution, normal distribution, geometric distribution, Poisson distribution, are computable (with computable parameters). Hence the change from distribution-free learning to computable-distribution-free learning is not too restrictive. It turns out that there is a nice mathematical structure in our computable-distribution-free learning case. For example, we can prove completeness results in the sense that there is a single (universal) distribution $\mathbf{m}$ such that if a concept class is learnable under this *single* distribution, they it is learnable under *all* computable distributions. Formally,

Theorem 2 *A concept class C is polynomially learnable under the universal distribution $\mathbf{m}(x)$, iff it is polynomially learnable under each computable distribution P, provided the sample is drawn according to $\mathbf{m}$.*

See [12] for details. In the continuous case, we even have a stronger theorem without needing to sample according to the universal distributions.

Theorem 3 *A concept class C over a continuous sample space is learnable under $\mathbf{M}$ iff it is learnable under each computable measure.*

5 Can we abandon pumping lemmas?

In the current undergraduate formal language courses, it seems that the cumbersome pumping lemmas constitute an important part of the teaching. It may be argued that such lemmas not only obstructs students' ability of viewing the real substance of the proof, but also give them a bad habit (like what 'goto' did to FORTRAN). Further, the usual pumping lemmas do not hold conversely which adds more confusion. Often students need un-aesthetic add-ons like "marked pumping lemma".

It turns out that Kolmogorov complexity is just the right tool to characterize *all* regular languages. It simply makes our intuition of 'finite state'-ness of these languages rigorous and easy to apply.

Theorem 4 (Regular KC-Characterization) *Let $L \subseteq \Sigma^*$, $\chi = \chi_1\chi_2\ldots$ be the* characteristic sequence *of $L_x = \{y|xy \in L\}$. The following statements are equivalent.*
(i) L is regular.
(ii) $\exists c_L$, $\forall x \in \Sigma^*$, $\forall n$, $C(\chi_{1:n}|n) \leq c_L$, c_L depending only on L.
(iii) $\exists c_L$, $\forall x \in \Sigma^*$, $\forall n$, $C(\chi_{1:n}) \leq C(n) + c_L$, c_L depending only on L.
(iv) $\exists c_L$, $\forall x \in \Sigma^*$, $\forall n$, $C(\chi_{1:n}) \leq \log n + c_L$, c_L depending only on L.

PROOF. (i) → (ii) → (iii) → (iv) are simple. To show (iv) → (i), we need,

Claim 5 For each constant c there are only finitely many one-way infinite binary strings ω such that, for all n, $C(\omega_{1:n}) \leq \log n + c$.

D.W. Loveland [14] credits the following result to A. Meyer: For each constant c there are only finitely many $\omega \in \{0,1\}^\infty$ with $C(\omega_{1:n}|n) \leq c$ for all n and each such ω is a recursive real. G.J. Chaitin [3] improved this to $C(\omega_{1:n}) \leq \log n + c$.

By (iv) and Claim 5, there are only finitely many distinct χ's associated with the x's in Σ^*. Define the right-invariant equivalence relation $\sim$ by $x \sim x'$ if $\chi = \chi'$. This relation induces a partition of Σ^* in equivalence classes $[x] = \{y : y \sim x\}$. Thus there are only finitely many $[x]$'s, which implies that L is regular by the Myhill-Nerode theorem: define a finite automaton using one state for each equivalent class, and define transition functions accordingly. □

See [13] for a complete proof, and for CFL's. As a corollary, we have

Lemma 6 (KC-Regularity) *Let L be regular and $L_x = \{y : xy \in L\}$. For each x, if y is the nth string is in L_x, then $C(y) \leq C(n) + c$, for some constant c.*

EXAMPLE 5.1 Consider $\{0^k 1^k : k \geq 1\}$. Set $x = 0^n$ with $C(n) \geq \log n$. Then lexicographically first string in L_x is 1^n, but $C(1^n) = \Omega(\log n)$. Thus KC-Regularity Lemma implies that $\{0^k 1^k : k \geq 1\}$ is not regular.

Comment. Intuitively, when a finite automaton reads 1^n, it has to remember n, but $C(n) \geq \log n$, a *finite* automaton cannot encode these $\log n$ bits.

EXAMPLE 5.2 Prove that $L = \{xx^R w : x, w \in \{0,1\}^* - \{\epsilon\}\}$ is not regular. Set $x = (01)^n$, where $C(n) \geq \log n$. Then, the lexicographically first word in L_x is $y = (10)^n 0$. Hence, $C(y) = \Omega(\log n)$, contradicting the KC-Regularity Lemma.

6 Loschmidt's Paradox

The second law of thermodynamics says that in any thermodynamic process that proceeds from one equilibrium state to another, the entropy of the system + environment either increase or remains unchanged. Fundamentally, this law says that any system, without external influence, goes to maximum disorder.

In 1872, Ludwig Boltzmann (1844-1906), lectured on the foundation of modern statistical thermodynamics. When he mentioned his interpretation of the second law, the physicist Joseph Loschmidt rose to protest. Loschmidt said that the laws governing the motions of all particles are symmetric with respect to time. Thus any system that goes from order to chaos could be made orderly once again by reversing the momentum of each particle. This will not affect the total kinetic energy of the system. In defiance Boltzmann pointed his finger at Loschmidt and said, "You reverse the momenta." [2]. Loschmidt in fact raised the following question: if the system is deterministic, how could its entropy be increased? In fact, Shannon entropy ($H = -\sum_s p_s \log p_s$) or, equivalently, Boltzmann-Gibbs entropy ($S = k \log W$) for a deterministic system cannot increase, because the number of possible states (W) does not increase in a deterministic system by Liouville's Theorem.

With Shannon's information theory, one cannot directly define the entropy of a *single* microscopic state. Sometimes such a definition is actually desired. For instance, in the high temperature superconductivity research, some material like CuO_2 loses magnetic moment below some critical temperature. In such state, the nuclear spins in CuO_2 all line up as in Figure 1.

$$
\begin{array}{l}
\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow \ \ldots\ldots \\
\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow \ \ldots\ldots \\
\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow \ \ldots\ldots \\
\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow \ \ldots\ldots
\end{array}
$$

Figure 1: Atomic spin in CuO_2 at low temperature

The state represented by Figure 1 is considered to have low entropy, or zero entropy. It is necessary to define entropy for such particular state, and not involve a whole ensemble as information theoretic entropy. It is not convenient to use the Shannon entropy in this situation since we might also like to know the entropy of the state with all arrows up, as in Ferromagnetism, or all arrows down. Or two arrows up and two arrows down alternatively, *etc.*

The above discussion naturally leads to the following new definitions of entropy for physical systems: *algorithmic entropy* and *physical entropy.* As we will demonstrate, the new definitions are not only quantitatively and philosophically correct, but also more direct, intuitive and elementary than the Shannon entropy.

Definition 1 The *algorithmic entropy* of a microscopic state of a system is the Kolmogorov complexity of that state.

The difficulties we had with Shannon entropy disappear in the new definition. Algorithmic entropy of a system can increase, even for a deterministic system, simply because the system evolves over time from regular initial states to more random states. Thus the second law is naturally and rigorously explained with algorithmic entropy. Since most states are random, a random system or even a deterministic system approaches and stays on average in states with maximum entropy. Note that second law was never meant to be true all of the time. The upshot is that the probability of the second law not to hold is very very low.

EXAMPLE 6.1 Regular microscopic states now automatically have well-defined low algorithmic entropies. The state of CuO_2 in Figure 1 naturally has low algorithmic entropy since it can be described by a trivial program of a few bits:

$$\textbf{repeat forever}: \ \textbf{print} \ \uparrow; \ \textbf{print} \ \downarrow.$$

Other easily describable states, such as with all arrows up or alternatively with two arrows up and two arrows down, also naturally have constant algorithmic entropy.

EXAMPLE 6.2 [This example is taken from D. Halliday and R. Resnick, *Fundamentals of Physics*, 3rd extended edition, 1988, page 526] Algorithmic entropy considerations lie at the heart of *adiabatic demagnetization*, an important method that has been used with great success to achieve record low temperatures (near zero Kelvin). In this method, a sample such as a chrome-alum salt (whose atoms may be considered as tiny magnets) is placed in an insulating enclosure at the lowest attainable temperature. A strong magnetic field is applied by an external magnet so that the tiny atomic magnets line up, forming a very ordered state, as Figure 2 (a)

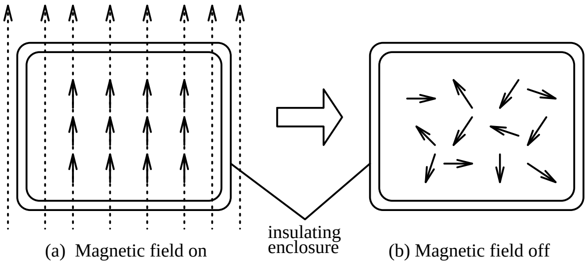

Figure 2: Adiabatic demagnetization to achieve low temperature

shows. Then the magnet is removed so that its field is no longer present. By thermal agitation, the atomic magnets now assume random orientations, as in Figure 2 (b). Kolmogorov complexity, *i.e., algorithmic entropy*, associated with the atomic alignments has clearly increased.

Since the system is adiabatically isolated, no heat can leave or enter. Since the process of removing the magnetic field is almost reversible, there can be no change in entropy in this thermally isolated *reversible* process. Thus the increase of entropy associated with the randomizing of the directions of the atomic magnets must be compensated with the spontaneous lowering of the temperature of the specimen, decreasing the entropy due to thermal agitation by the same amount.

We further justify Definition 1. Given a thermodynamic ensemble $\mathcal{E}$, we prove that Shannon entropy and algorithmic entropy are quantitatively the same.

Theorem 7 *Let $P(s)$ be the probability that the system is in state $s \in \mathcal{E}$. If probabilities $P(s)$ can be recursively generated, then, up to a constant additive factor,*

$$H(P) = \sum_{s \in \mathcal{E}} P(s) K(s), \tag{6}$$

where $H(P) = -\sum_{s \in \mathcal{E}} P(s) \log P(s)$ is the Shannon entropy.

PROOF. By the Noiseless Coding Theorem,

$$H(P) \leq \sum_{s \in \mathcal{E}} P(s) K(s).$$

By a constructive version of the Shannon-Fano code, $K(s) \leq -\log P(s) + O(1)$. It follows that

$$\sum_{s \in \mathcal{E}} P(s) K(s) \leq H(P) + O(1).$$

[13] contains all the complex details. □

Equation 6 shows that algorithmic entropy quantitatively approximates the old Shannon entropy. Thus algorithmic entropy naturally inherits the successes and results of the Shannon entropy. As an example, we prove one such result, the Sackur-Tetrode equation [24].

Ideal gas or Boltzmann gas is a convenient tool to study statistical thermodynamics. Consider a 3-dimensional container with N identical (ideal) gas molecules. Each molecule is considered as an elastic ball with no internal freedom. Imagine we divide the entire container volume into a 3-dimensional array of small cells, each capable of containing one molecule. At any fixed time, a specific cell contains zero or one molecule. To specify N molecules, we use a *phase-space* of $6N$ dimensions. A particular microscopic state corresponds to just one point $P = (p_1, \cdots, p_{3N}, q_1, \cdots, q_{3N})$ in phase-space, where p_i's are coordinates of the positions of N molecules and q_i's specify their momenta. This is called a Hamiltonian system. (In Boltzmann entropy, $S = k \log W$, W is the number of points in our phase-space corresponding to the macroscopic state of entropy S. Liouville's Theorem says any region in the phase-space does not change in volume as the system evolves with time. Thus as a system evolves, W, therefore S, stays constant. S correctly specifies the equilibrium entropy, but it lacks the ability of specifying the dynamic changes of the system, as prescribed by the second law.)

Theorem 8 *(Sackur-Tetrode Equation) Let V be the volume of the container, and ΔV be the volume of each cell. Let Δ_q be the basic unit we record momentum q. Then the entropy of a typical microscopic state of N ideal molecules is given by the following formula:*

$$S \leq N[\log \frac{V}{N\Delta V} + \frac{3}{2} \log \frac{mkT}{(\Delta_q)^2}] + O(1).$$

PROOF. We design a program of above size to specify any given microscopic state, assuming N is given. The positions of molecules can be specified as follows: There are $V/\Delta V$ cells and N molecules. To specify the distances of the molecules, we need $\sum_{i=1}^{N} d_i$ bits, where $\sum d_i = V/\Delta V$. Maximizing this, we conclude that at most $N \log \frac{V}{N\Delta V}$ bits are needed to describe the positions of the N molecules.

The expected value of each component of momentum q of a molecule is $(mkT)^{1/2}$, thus to specify the momentum of such a molecule requires $3 \log \frac{(mkT)^{1/2}}{\Delta_q}$ bits. In total, we need at most $\frac{3}{2} N \log \frac{mkT}{(\Delta_q)^2}$ bits to specify the momenta of N molecules. □

Algorithmic entropy is a definition from the system point of view. It defines the complexity of a system at a particular time. If we look at a system from an observer's angle, W.H. Zurek [24] defined the following 'physical entropy'.

Definition 2 Physical entropy of a system, given observed data d of the system, is $H_d + K(d)$, where $K(d)$ is the Kolmogorov complexity of the observed data d, and H_d is the conditional Shannon entropy or our "ignorance" of the system given d.

Physical entropy reflects the fact that measurements can increase our knowledge about a system. If the system is in a regular state, physical entropy can decrease as we make more and more measurements. In the beginning, we have no knowledge about the state of system, therefore the physical entropy reduces to Shannon entropy and is maximized reflecting our total ignorance. As we make more measurements, if the state is not regular, then we cannot achieve compression and the physical entropy remains high. However, if the system is in a regular state, then measurements

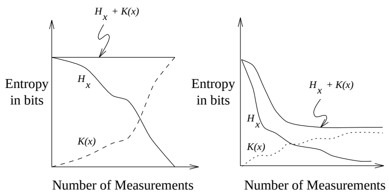

Figure 3: Physical Entropy

increases our knowledge about the system and we might be able to compress the data, and the system becomes less unknown and hence physical entropy deminishes. The two pictures in Figure 3 describe these situations.

Notice that the physical entropy in Definition 2 cannot increase if we make no measurements. However, physical entropy can increase if we made sufficient measurements at different time steps as the system evolves and becomes more and more disordered. It is important to distinguish the increase of $K(d)$ with measurements in Figure 3 and the increase of algorithmic entropy with time. In the former case, we have more and more data at hand to describe, therefore $K(d)$ grows until it describes the systems totally. During this process, the system is assumed to be static with the arrow of time frozen. In the latter case, the system evolves to become more and more chaotic with time, according to the second law. Thus if we make sufficient amount of measurements to the system at some later time step, $K(d)$ will be greater, reflecting the fact that the system is now more disordered, or, has greater algorithmic entropy. This increase will stop at the equilibrium of the system. Zurek [25] shows how to use physical entropy to explain Maxwell's demon.

7 Why is our world compressible?

P.C.W. Davies [4] asked the interesting question why the world is compressible. Science, in a sense, may be considered as compression of experimental data. Compression means knowing and comprehending. Why is mathematics compressible to a few axioms? Why is physics (approximately) compressible to a few laws? Why are we humans compressible to a string over alphabet $\{A, C, G, T\}$? Why is the universe compressible, and hence comprehensible, at all?

E. Wigner [22] called this phenomenon "the unreasonable effectiveness" of mathematics in the natural sciences. We are so used to our environment, that above questions probably have never puzzled most of us. If you do not realize the subtleties of these questions, listen to R.W. Hamming [*Coding and Information Theory*, Prentice-Hall, 1987]:

> I have tried, with little success, to get some of my friends to understand my amazement that the abstraction of integers for counting is both possible and useful. Is it not remarkable that 6 sheep plus 7 sheep makes 13 sheep; that 6 stones plus 7 stones make 13 stones? Is it not a miracle that the universe is so constructed that such a simple abstraction as a number is possible? To me this is one of the strongest examples of the unreasonable effectiveness of mathematics. Indeed, I find it both strange and unexplainable.

Is it unreasonable to imagine a total chaotic universe? As we know, a randomly given string (or universe) is most likely to be not algorithmically compressible. How come ours happens to be algorithmically compressible, hence knowable? More remarkably, our universe is not only compressible, it is indeed very *feasibly* compressible, since mankind is good at discovering nature's algorithms. It took only one man and one apple and 13 years to invent Newton mechanics (and calculus); it took only one man and his life time to discover evolution theory; it took only one man and a few years to understand relativity; it took only a few more men and a few more years to formulate quantum mechanics. Yet, we obtain so much from so little: we sent men to the moon with a few laws of Newton; we release gigantic nuclear energy with Einstein's $E = mc^2$; our TV's, radios, X ray, and police radar all depend on just 4 lines of Maxwell equations; and with quantum mechanics and superstring theory, we hope that they explain everything on earth, and in heaven.

Let us look at things from a different perspective. Life, in some sense, also evolves in the direction of minimizing its 'programs', and indeed this turns out to be not only possible but also extremely effective. A DNA molecule can be regarded as a long character string over $\{A, C, G, T\}$. In order to encode proteins, it needs to encode 20 amino acids, plus a 'begin' and an 'end' command to signify the start and the end of an encoding of a protein. Thus to encode each of the 22 different objects, one needs at least three characters which give 64 possible combinations, while two characters only give 16 combinations. Using more than three characters is obviously redundant. Remarkably, indeed Nature uses precisely three characters to encode an amino acid or a 'begin' or an 'end' command, not more and not less. (Note, we do not claim that such sequences have maximum Kolmogorov complexity, in fact they do not.) More remarkably, Nature also dynamically minimizes its programs through Darwinian selection. To demonstrate this point, we describe an experiment performed by S. Spiegelman [19], discussed in D.K. Kondepudi [9]. A Q_β virus replicates using the resources of the cell it infects. Since intra-cellular environment is usually complex and malicious, for successful replication, this virus carries in its RNA the algorithm to synthesize proteins that form a protective coat. Such RNA has about 4500 units. Spiegelman placed such RNA of 4500 units in a friendly environment conducive to replication and let it evolve. Soon, mutations with smaller number of units that could replicate faster arose and replaced the original RNA. This process continued until the RNA was reduced to about 220 units. This is an application of the MDL principle above.

The current laws of physics of this universe only started to hold at approximately 10^{-4}s of the big bang. According to the best cosmological theories, the universe began in an exceedingly simple state. Is this the reason that our current universe is highly compressible? Using a binary string to encode the universe, we can ask the question: what is the Kolmogorov complexity of the world? C.H. Woo [23] argues that it maybe premature to claim that the world is indeed highly (algorithmically) compressible. Even the universe would be maximally random, then our part of the universe could be regular and compressible, since every long string has short compressible substrings. Either way, this will stay unknown, since a random string cannot be proven to be random effectively. Relying on quantum mechanics, R. Penrose in his popular book [*Emperor's New Mind*, Oxford University Press, 1989] has conjectured that the human brain has capabilities superior to that of a Turing machine. Does God really permit us to learn more about Him via proofs imagined by our minds than proofs listed by our hands?

Among other interesting philosophical issues relating Kolmogorov complexity to physics are: Thermodynamics of computing and Maxwell's demon (can one make a perpetual machine of second kind?) [1, 24], chaos theory (can a deterministic system be chaotic?) [6], visual distance, *etc.* The interesting book [25] contains a wealth of related papers. [13] tries to present a complete treatment. [21] contains five surveys on applications of Kolmogorov complexity to structural complexity theory.

References

[1] C.H. Bennett. The thermodynamics of computation—a review. *Int'l J. Theoret. Physics*, 21:905–940, 1982.

[2] R.G. Brewer and E.L. Hahn. Atomic memory. *Scientific American*, 251(6):50–57, 1984.

[3] G.J. Chaitin. Information-theoretic characterizations of recursive infinite strings. *Theoret. Comput. Sci.*, 2:45–48, 1976.

[4] P.C.W. Davies. Why is the physical world so comprehensible? In W.H. Zurek, editor, *Complexity, entropy and the physics of information*, pages 61–70. Addison-Wesley, 1991.

[5] P. Erdös and A. Rényi. On two problems of information theory. *Publ. Hungar. Ac. Sci.*, 8:241–254, 1963.

[6] J. Ford. Chaos: solving the unsolvable, predicting the unpredictable. In M.F. Barnsley and S.G. Demko, editors, *Chaotic dynamics and fractals*. Academic Press, 1986.

[7] N. Goodman. *Fact, Fiction, and Forecast*. Harvard University Press, 1955.

[8] A.N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Math. Surveys*, 38(4):29–40, 1983.

[9] D.K. Kondepudi. Non-equilibrium polymers, entropy, and algorithmic information. In W.H. Zurek, editor, *Complexity, entropy and the physics of information*, pages 199–206. Addison-Wesley, 1991.

[10] M. Li and P.M.B. Vitányi. Two decades of applied Kolmogorov complexity: In memoriam A.N. Kolmogorov 1903 - 1987. In *Proc. 3rd IEEE Conference on Structure in Complexity Theory*, pages 80–101, 1988.

[11] M. Li and P.M.B. Vitányi. Kolmogorov complexity and its applications. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, chapter 4, pages 187–254. Elsevier and MIT Press, 1990.

[12] M. Li and P.M.B. Vitányi. A theory of learning simple concepts under simple distributions. *SIAM J. Computing*, 20(5):915–935, 1991.

[13] M. Li and P.M.B. Vitányi. *An introduction to Kolmogorov complexity and its applications.* Springer-Verlag, New York. 1993.

[14] D.W. Loveland. A variant of the Kolmogorov concept of complexity. *Inform. Contr.*, 15:510–526, 1969.

[15] L. Moser. The second moment method in combinatorial analysis. In *Combinatorial Structures and their applications*, pages 283–384. Gordon and Breach, New York, 1970.

[16] N. Pippenger. An information-theoretic method in combinatorial theory. *J. Comb. Theory, Ser. A*, 23:99–104, 1977.

[17] J. Rissanen. Modeling by the shortest data description. *Automatica-J.IFAC*, 14:465–471, 1978.

[18] R.J. Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Inform. Theory*, IT-24:422–432, 1978.

[19] S. Spiegelman. An in vitro analysis of a replicating molecule. *American Scientist*, 55(3):221–264, 1967.

[20] L.G. Valiant. A theory of the learnable. *Comm. Assoc. Comput. Mach*, 27:1134–1142, 1984.

[21] O. Watanabe, editor. *Kolmogorov complexity and its relation to computational complexity theory.* To appear.

[22] E. Wigner. The unreasonable effectiveness of mathematics in natural sciences. *Comm. Pure Appl. Math*, 13:1, 1960.

[23] C.H. Woo. Laws and boundary conditions. In W.H. Zurek, editor, *Complexity, entropy and the physics of information*, pages 127–135. Addison-Wesley, 1991.

[24] W.H. Zurek. Algorithmic randomness and physical entropy. *Physical Review, series A*, 40(8):4731–4751, 1989.

[25] W.H. Zurek, editor. *Complexity, entropy and the physics of information.* Addison-Wesley, 1991.