Kolmogorov Complexity: Sources, Theory and Applications

ALEXANDER GAMMERMAN AND VLADIMIR VOVK

Computer Learning Research Centre and Department of Computer Science,
Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK
Email: alex@dcs.rhbnc.ac.uk

We briefly discuss the origins, main ideas and principal applications of the theory of Kolmogorov complexity.

1. UNIVERSALITY

The theory of Kolmogorov complexity is based on the discovery, by Alan Turing in 1936, of the universal Turing machine. After proposing the Turing machine as an explanation of the notion of a computing machine, Turing found that there exists one Turing machine which can simulate any other Turing machine.

Complexity, according to Kolmogorov, can be measured by the length of the shortest program for a universal Turing machine that correctly reproduces the observed data. It has been shown that, although there are many universal Turing machines (and therefore many possible 'shortest' programs), the corresponding complexities differ by at most an additive constant.

The main thrust of the theory of Kolmogorov complexity is its 'universality'; it strives to construct universal learning methods based on universal coding methods. This approach was originated by Solomonoff and made more appealing to mathematicians by Kolmogorov. Typically these universal methods will be computable only in some weak sense. In applications, therefore, we can only hope to approximate Kolmogorov complexity and related notions (such as randomness deficiency and algorithmic information mentioned below). This special issue contains both material on non-computable aspects of Kolmogorov complexity and material on many fascinating applications based on different ways of approximating Kolmogorov complexity.

2. BEGINNINGS

As we have already mentioned, the two main originators of the theory of Kolmogorov complexity were Ray Solomonoff (born 1926) and Andrei Nikolaevich Kolmogorov (1903–1987). The motivations behind their work were completely different; Solomonoff was interested in inductive inference and artificial intelligence and Kolmogorov was interested in the foundations of probability theory and, also, of information theory. They arrived, nevertheless, at the same mathematical notion, which is now known as Kolmogorov complexity.

In 1964 Solomonoff published his model of inductive inference. He argued that any inference problem can be presented as a problem of extrapolating a very long sequence of binary symbols; 'given a very long sequence, represented by T, what is the probability that it will be followed by a ... sequence A?'.

Solomonoff assumed that all the information required for the induction was contained in the string itself and in the spirit of the Bayesian approach, assigned the *a priori* probability (or *algorithmic probability measure*)

$$P(x) = \sum_{p} 2^{-l(p)}$$

to a binary string x, where the sum is taken over all programs p for generating x and $l(p)$ stands for the length of the string p.

The problem with this approach is that $\sum_x P(x)$ diverges; moreover, $P(x) = \infty$ for each x. The later attempts to fix this problem by Ray Solomonoff resulted in the algorithmic probability measure becoming non-computable even in a weak sense of 'semicomputability from below' (for the precise definition, see V'yugin's paper in this issue). However, Solomonoff's idea of measuring the complexity of an object using universal Turing machines and his definition of algorithmic probability measure were major developments in the history of this approach.

In 1965 Kolmogorov published a paper on the algorithmic approach to information theory where he defined the complexity of a finite object and also showed how to measure the amount of mutual information in one (finite) object about another. Intuitively, it was clear to everyone that there are 'simple' objects and 'complex' ones; the problem was the diversity of the ways of describing objects: an object can have a 'simple' description (i.e. short) in one language but not in another. The discovery made by Solomonoff and Kolmogorov was that with the help of the theory of algorithms it is possible to restrict this arbitrariness and define complexity as an invariant concept.

For Kolmogorov the main motivation behind his work on algorithmic complexity was perhaps his desire to formalize the notion of a random sequence. It is easy to see that, if x is an element of a 'simple' (in the sense of Kolmogorov complexity) finite set A, the Kolmogorov complexity $K(x)$ of x cannot be much greater than the binary logarithm $\log |A|$ of the size of A. This simple upper bound on $K(x)$

allowed Kolmogorov to define the notion of randomness; x is *random* in A if $K(x)$ is close to its upper bound $\log |A|$.

Kolmogorov's ideas of randomness were almost immediately developed further by his students and followers, notably by Per Martin-Löf [1] and Leonid Levin. In particular, Martin-Löf extended Kolmogorov's notion of randomness from sets to probability distributions; it became clear that if P is a simple probability distribution in a simple finite set A, we can define an element $x \in A$ to be random if its Kolmogorov complexity is close to its upper bound $-\log P(x)$. Kolmogorov's definition is a special case of this latter definition corresponding to P taken to be the uniform distribution in A.

At about the same time, in 1966, Gregory Chaitin published a paper [2] on the complexity of algorithms and later, in 1969 [3], he developed his ideas further in studying infinite random sequences; as a byproduct, in the 1969 paper he arrived at the notion of Kolmogorov complexity. Among Chaitin's most well-known achievements are an extension of Gödel's incompleteness theorem and a study of the limitations of the axiomatic method.

The history of developing the notion of Kolmogorov complexity is described in more detail in Li and Vitányi [4].

3. THEORY

Traditionally the theory of Kolmogorov complexity is divided into three parts: complexity proper; randomness; information. The term we use in this special issue, the theory of Kolmogorov complexity, emphasizes complexity. Two other widely used terms are 'algorithmic information theory' (emphasizing the information part) and 'algorithmic probability theory' (emphasizing the randomness part). In this section we briefly review these three parts.

3.1. Complexity

One the most important developments in the theory of Kolmogorov complexity was the introduction of the notion of prefix complexity by Levin [5] (see also Gács [6]) and, independently of Levin, by Chaitin [7]. Prefix complexity has been used to solve some technical problems in algorithmic information theory (such as the lack of symmetry of the algorithmic information; see below) and the theory of algorithmic randomness.

Besides prefix complexity, several other complexities were introduced, such as Loveland's complexity of resolution and Levin's monotonic complexity (also independently defined by Schnorr). A general framework in which different complexity measures can be described in a uniform fashion was suggested by Alexander Shen [8]. All four variants of Kolmogorov complexity coincide to within a logarithmic term.

One especially useful variant of Kolmogorov complexity (lying outside Shen's [8] classification) is the minus logarithm of the *a priori* semi-measure, as defined by Solomonoff (see above) and, in a more satisfactory way, by Zvonkin and Levin [9]. This variant also coincides with the other variants to within a logarithmic term.

Kolmogorov complexity has found important practical applications in the form of Rissanen's minimum description length (MDL) principle and Wallace's minimum message length (MML) principle; see Section 4 below.

3.2. Randomness

As we discussed above, randomness essentially means the closeness of the Kolmogorov complexity to some upper bound. In [10] and [11] Kolmogorov sketched a programme of using his notion of randomness for establishing closer connections between the mathematical theory of probability and the applications of probability. However, the fate of this program was different from that of his earlier proposal, the universally accepted axioms of probability theory [12]. The theory of randomness was mainly developed in the case of infinite sequences; whereas Kolmogorov found the study of infinite sequences interesting, it is clear that any connection with reality is lost when attention is switched to the empirically non-existent infinite sequences (cf. Kolmogorov [11], the first paragraph of Section 6).

It is clear why infinite sequences were more appealing to mathematicians: we can divide all infinite sequences into two well-defined classes, the random sequences and the non-random sequences. In the case of finite sequences, typically all sequences will be random and mathematical results have to be stated in terms of degree of non-randomness or, technically, *randomness deficiency*. Since randomness deficiency is defined only to within an additive constant, such results often look less beautiful, from the purely mathematical point of view, than the results about infinite sequences.

The lack of practical applications is the most disappointing side of the algorithmic theory of randomness; this explains why the papers in this special issue are mainly devoted to the theory of Kolmogorov complexity proper (see the classification in the beginning of this section). The situation is changing now: work on applying the algorithmic theory of randomness to practical problems of computer learning is under way in the Computer Learning Research Centre at Royal Holloway, University of London; the reader can consult the web page `http://www.clrc.rhbnc.ac.uk`.

3.3. Information

Besides Kolmogorov's life-long interest in the foundations of probability theory, another motivation for his work on algorithmic complexity was the then new ideas of information theory. In [10] he defined the information in an object x about another object y as

$$I(x : y) = K(y) - K(y \mid x).$$

The main problem, noticed by Kolmogorov [10], with this definition of information is that some fundamental properties of the standard probabilistic notion of information, such as symmetry, cease to be true (even to within an additive constant) for the algorithmic information $I(x : y)$. A solution to this problem was suggested by Levin, Gács and Chaitin: it is sufficient to define

$$I(x : y) = K(y) - K(y \mid x, K(x)),$$

where K is the prefix complexity.

This third part of the theory of Kolmogorov complexity seems to be less developed than the first two parts. A possible explanation is that the information $I(x : y)$ is not computable even in the weak sense of semi-computability from below or above: both $K(y)$ and $K(y \mid x)$ are semi-computable from above, but their difference $I(x : y)$ is only computable in the limit.

3.4. Further reading

The first and still very useful review of the theory of Kolmogorov complexity is Zvonkin and Levin [9]. An excellent review first published in Russian in 1981 was V'yugin's paper [13]. The encyclopedia of Kolmogorov complexity is Li and Vitányi [4]. We also recommend the chapters devoted to Kolmogorov complexity and randomness in Uspensky and Semenov [14]. Several papers in this issue, especially V'yugin's paper, give reviews of different parts of the theory of Kolmogorov complexity.

4. APPLICATIONS

Probably the bulk of applications of Kolmogorov complexity is currently done in the framework of the MDL principle (Rissanen [15, 16, 17, 18]) and MML principle (Wallace and Boulton [19], Wallace and Freeman [20]). In this section we will only describe the main idea of the simplest versions of these two principles, as applied to the problem of model selection. In our description we will emphasize the role of Kolmogorov complexity.

Suppose we have some 'statistical model' $\{P_\theta\}$, which is a family of probability distributions P_θ, θ ranging over some parameter space Θ. In the most interesting and difficult cases the parameter space consists of a series of subspaces

$$\Theta = \Theta_1 \cup \Theta_2 \cup \ldots;$$

e.g. Θ might be the set of all polynomials (to be fitted to the data) and Θ_n, $n = 1, 2, \ldots$, might be the set of polynomials of degree n. For such 'big' models the usual methods of estimating θ given data x, such as the maximum likelihood estimator, do not work: if we take θ for which $P_\theta(x)$ (or the corresponding density if P_θ are continuous) attains its maximum, we will grossly overfit the data; taking the degree of the polynomial big enough, we will be able to fit the data precisely, but the performance on new points will typically be very poor.

It is clear that we need to somehow 'regularize' our estimate θ. The MDL and MML principles give very natural ways of regularization, in a way extending Occam's razor. For simplicity we will assume that the parameter space Θ and the set of all possible values for the data x are countable (e.g. we can assume that θ and x are measured with finite accuracy). We assume that the model $\{P_\theta\}$ is simple (this is always satisfied in applications).

In the case where $\{P_\theta\} = \{P\}$ consists of only one distribution P, we know (see the discussion of randomness in the previous section) that

$$K(x) \leq -\log P(x) + C; \qquad (1)$$

here and below K is the prefix complexity and C stands for a constant (which can change from one formula to another). Intuitively, when given a probability distribution P we can find a code (e.g. the Shannon–Fano code) which codes every x using at most $-\log P(x) + C$ bits; therefore, the Kolmogorov complexity of x, being defined in terms of the universal code, cannot exceed $-\log P(x) + C$.

Inequality (1) can be extended to general $\{P_\theta\}$ as follows: for all θ and x,

$$K(x) \leq -\log P_\theta(x) + K(\theta) + C. \qquad (2)$$

Intuitively, we can encode x in two steps: first we encode some θ and then we encode x using e.g. the Shannon–Fano code constructed from the probability distribution P_θ.

The simplest versions of the MDL and MML principles can be regarded as Kolmogorov complexity approximation principles: when given data x we choose θ which provides a minimum to an upper bound, such as (2), on $K(x)$. Notice that we obtain the maximum likelihood estimator by minimizing the right-hand side of (2) and ignoring the addend $K(\theta)$. In general, to minimize the right-hand side of (2) we have to balance the log-likelihood $-\log P(x)$ against the complexity $K(\theta)$ of the 'hypothesis' θ.

The MDL and MML principles use different strategies for approximating $K(x)$. It is clear that we cannot use inequality (2) directly since $K(\theta)$ is not computable. In the MDL principle one finds as good as possible an approximation (necessarily from above) to $K(\theta)$ (using perhaps Rissanen's [16] universal prior for integers). If we have a good upper bound $K^*(\theta)$ for $K(\theta)$, the MDL principle (in its simplest form) recommends choosing θ by minimizing the expression

$$-\log P_\theta(x) + K^*(\theta). \qquad (3)$$

The MML principle takes a Bayesian stance, assuming a pre-specified prior distribution on the parameter set. With this prior distribution we can find the expected value of (3); instead of using some 'universal' K^* in (3), the 'strict' MML principle recommends using the K^* which provides the minimum of the expected value of (3).

In this section we have only discussed the simplest versions of MDL and MML; actually both principles have generated a lot of research and there are several variants of them. Some of these variants can be interpreted as using more accurate approximations of Kolmogorov complexity.

5. THIS ISSUE

This special issue opens with a paper by Ray Solomonoff ('Two kinds of probabilistic induction'), one of the founders of the theory of Kolmogorov complexity. In that paper the theory of algorithmic probability (one more term for what we call the theory of Kolmogorov complexity) is extended to problems of induction which are not explicitly sequential.

Jorma Rissanen, the author of the MDL principle, shows how the MDL principle is applied to hypothesis selection and testing.

Chris Wallace, the author of the MML principle, and David Dowe look at a parallel between a version of the Kolmogorov model and the MML approach to inference.

Leonid Levin, one of the creators of the modern theory of Kolmogorov complexity, presents an outline of some robust measures of information and suggests various applications of those measures.

Tao Jiang, Ming Li and Paul Vitányi demonstrate the power of the incompressibility method, one of the most impressive applications of Kolmogorov complexity.

Vladimir V'yugin presents the mathematical review of the theory and some applications of Kolmogorov complexity. It includes both well-known results and newer results which have not appeared in the review literature so far.

Vladimir Vovk and Alexander Gammerman propose to generalize the MDL and MML principles to a general inductive 'complexity approximation principle'.

Besides the eight 'main' papers, we decided to include in this special issue a discussion section on the MDL and MML inductive principles, which play such an important role in applications. There was such a discussion in 1987 [17, 20], but since then there have been many new developments.

The discussion is opened by Phil Dawid, who also opened the Royal Statistical Society discussion in 1987. Then Jorma Rissanen discusses the MML principle and Chris Wallace and David Dowe discuss the MDL principle in the framework of Kolmogorov complexity. The contribution to the discussion presented by Bernard Clarke puts MDL and MML in a wider context of model selection principles (such as the Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC), etc). Alexander Shen discusses the connections between Kolmogorov complexity and statistical analysis in general and the role of the MML and MDL principles in this context.

The last part of the issue consists of the rejoinders by the authors of the MDL and MML principles to the comments and criticism made by the contributors to the discussion part.

ACKNOWLEDGEMENTS

This work was partially supported by EPSRC through grants GR/L35812 ('Support vector and Bayesian learning algorithms'), GR/M14937 ('Predictive complexity: recursion-theoretic variants'), and GR/M16856 ('Comparison of support vector machine and minimum message length methods for induction and prediction').

REFERENCES

[1] Martin-Löf, P. (1966) The definition of random sequences. *Information and Control*, **9**, 602–619.

[2] Chaitin, G. J. (1966) On the length of programs for computing finite binary sequences. *J. Assoc. Comput. Mach.*, **13**, 547–569.

[3] Chaitin, G. J. (1969) On the length of programs for computing finite binary sequences: statistical considerations. *J. Assoc. Comput. Mach.*, **16**, 145–159.

[4] Li, M. and Vitányi, P. (1997) *An Introduction to Kolmogorov Complexity and Its Applications* (2nd edn). Springer, New York.

[5] Levin, L. A. (1974) Laws of information conservation (nongrowth) and aspects of the foundations of probability theory. *Problems Inform. Transmission*, **10**, 206–210.

[6] Gács, P. (1974) On the symmetry of algorithmic information. *Sov. Math.–Dokl.*, **15**, 1477–1480.

[7] Chaitin, G. J. (1975) A theory of program size formally identical to information theory. *J. Assoc. Comput. Mach.*, **22**, 329–340.

[8] Shen, A. (1984) Algorithmic variants of the notion of entropy. *Sov. Math.–Dokl.*, **29**, 569–573.

[9] Zvonkin, A. K. and Levin, L. A. (1970) The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Math. Surveys*, **25**, 83–124.

[10] Kolmogorov, A. N. (1968) Logical basis for information theory and probability theory. *IEEE Trans. Inform. Theory*, **IT-14**, 662–664.

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

[12] Kolmogorov, A. N. (1933) *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin. Published in English as *Foundations of the Theory of Probability*. Chelsea, New York. 1st edn, 1950; 2nd edn, 1956.

[13] V'yugin, V. V. (1994) Algorithmic entropy (complexity) of finite objects and its applications to defining randomness and amount of information. *Selecta Mathematica Sovietica*, **13**, 357–389.

[14] Uspensky, V. A. and Semenov, A. L. (1993) *Algorithms: Main Ideas and Applications*. Kluwer, Dordrecht.

[15] Rissanen, J. (1978) Modeling by the shortest data description. *Automatica*, **14**, 465–471.

[16] Rissanen, J. (1983) A universal prior for integers and estimation by minimum description length. *Ann. Statist.*, **11**, 416–431.

[17] Rissanen, J. (1987) Stochastic complexity (with discussion). *J. R. Statist. Soc.*, B, **49**, 223–239 and 252–265.

[18] Rissanen, J. (1989) *Stochastic Complexity and Statistical Inquiry*. World Scientific, Singapore.

[19] Wallace, C. S. and Boulton, D. M. (1968) An information measure for classification. *Comput. J.*, **11**, 185–195.

[20] Wallace, C. S. and Freeman, P. R. (1987) Estimation and inference by compact coding (with discussion). *J. R. Statist. Soc.* B, **49**, 240–265.