Lecture Notes in
Computer Science 623

W. Kuich (Ed.)

Automata, Languages and Programming

19th International Colloquium
Wien, Austria, July 1992
Proceedings

EATCS

Springer-Verlag

W. Kuich (Ed.)

Automata, Languages and Programming

19th International Colloquium 《计 纪会
Wien, Austria, July 13-17, 1992
Proceedings

EATCS

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Volume Editor

Werner Kuich
Abt. für Theoretische Informatik, Inst. für Algebra und Diskrete Mathematik
Technische Universität Wien, Wiedner Hauptstraße 8-10, A-1040 Wien, Austria

CR Subject Classification (1991): D.1, E.1, F, I.1, G.2, I.2.3, I.3.5

ISBN 3-540-55719-9 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-55719-9 Springer-Verlag New York Berlin Heidelberg

Printed in Germany

Typesetting: Camera ready by author/editor
Printing and binding: Druckhaus Beltz, Hemsbach/Bergstr.
45/3140-543210 - Printed on acid-free paper

PREFACE

ICALP 92 is the 19th International Colloquium on Automata, Languages, and Programming in a series of meetings sponsored by the European Association for Theoretical Computer Science (EATCS). It is a broadly based conference covering all aspects of Theoretical Computer Science including topics like: computability, automata, formal languages, term rewriting, analysis of algorithms, computational geometry, computational complexity, symbolic and algebraic computation, cryptography, data types and data structures, theory of data bases and knowledge bases, semantics of programming languages, program specification, transformation and verification, foundations of logic programming, theory of logical design and layout, parallel and distributed computation, theory of concurrency, theory of robotics.

ICALP 92 was held at Technische Universität Wien from July 13 to July 17, 1992. The Organizing Committee consisted of G. Baron, P. Kirschenhofer, W. Kuich (Chairman), H. Maurer, H. Prodinger, F. Urbanek.

Previous ICALPs were held in Madrid (1991), Warwick (1990), Stresa (1989), Tampere (1988), Karlsruhe (1987), Rennes (1986), Nafplion (1985), Antwerp (1984), Barcelona (1983), Aarhus (1982), Haifa (1981), Amsterdam (1980), Graz (1979), Udine (1978), Turku (1977), Edinburgh (1976), Saarbrücken (1974), Paris (1972). ICALP 93 will be held in Lund, Sweden, from July 5 to July 9, 1993.

The Programme Committee of ICALP 92 consisted of G. Ausiello, J. Berstel, R. V. Book, B. Buchberger, R. Cori, R. De Nicola, H. Edelsbrunner, S. Even, Ph. Flajolet, D. Harel, M. C. B. Hennessy, N. D. Jones, J. W. Klop, D. C. Kozen, W. Kuich (Chairman), B. Monien, E.-R. Olderog, B. Rovan, A. Salomaa, S. Skyum, A. Tarlecki. It has selected 52 papers from 190 papers submitted in response to the call for papers. These papers came from the following countries: Australia, Austria, Belgium, Canada, China, CSFR, Denmark, Finland, France, FRG, Greece, Hungary, India, Israel, Italy, Japan, Latvia, Lithuania, Netherlands, Poland, Russia, Singapore, Spain, Sweden, Switzerland, Taiwan, UK, Ukraine, USA. Each submitted paper has been evaluated by at least four members of the Programme Committee. The final selection was made during the selection meeting on February 1–2, 1992 in Wien, in which all Programme Committee members except S. Even and D. Harel could participate. Together with six invited presentations all 52 selected papers are contained in this volume. It is a pleasure for the conference chairman to thank the members of the Programme Committee for their evaluation of the papers, and the many referees who assisted in this process. The list of referees given in this volume is as complete as we could achieve, and we apologize for any possible omissions or errors.

The papers in this volume are printed in the order of presentation at ICALP 92 and thus grouped into sessions, most of which are thematic.

Wien, April 1992

Werner Kuich

LIST OF REFEREES

Aalbersberg H.
Aceto L.
Allouche J.-P.
Amtoft T.
Andersen N.
Ardeleanu E.
Arnold A.
Attiya H.
Atzeni P.
Ausiello G.
Autebert J.-M.
Avenhaus J.
Baaz M.
Baeten J.
Baron G.
Bar-On I.
Barringer H.
Bar-Yehuda R.
Bauer
Beauquier D.
Bednarczyk M.
Ben-David S.
Berstel J.
Bertoni A.
Best E.
Bialasik M.
Birk Y.
Boasson L.
Bodlaender H.
Boissonnat J.
Bol R. N.
Bolognesi T.
Bonuccelli M. A.
Book R. V.
Borzyszkowski A.
Boucheron St.
Bruyere V.
Buchberger B.
Buntrock G.
Capocelli R.
Cerone A.
Chazelle B.
Choffrut C.
Chor B.
Chretienne P.
Chytil M.
Collins G.
Cori R.
Courcelle B.
Crépeau C.
Crochemore M.
Czaja L.
Dallery Y.
Damgaard I.
D'Amore F.
Dauchet M.
De Boer F.
De Fraysseix H.
Delfinado C.
De'Lignoro U.
De Luca A.
De Nicola R.
Devillers O.
De Vink E.
De Vries F.-J.
Dey T.
Di Battista G.
Diekmann R.
Drabent W.
Dumas Ph.
Duris P.
Dybjer P.
Dybkjaer H.
Edelsbrunner H.
Engelfriet J.
Even S.
Facello M.
Feldmann R.
Feuerstein E.
Flajolet Ph.
Franciosa P. G.
Frandsen G. S.
Fricker C.
Friedman N.
Frougny Ch.
Gallier J.
Gambosi G.
Gastin P.
Gerth R.
Gnesi S.
Goldreich O.
Gomard C.
Goralcik P.
Gorrieri R.
Gouyou-Beauchamps D.
Habel A.
Habib M.
Hagerup T.
Han Chiang-Chih
Hannan A.
Harel D.
Harju T.
Henglein F.
Hennessy M. C. B.
Hertrampf U.
Hilken B.
Hong H.
Honkala J.
Hromkovic
Huizing C.
Hungar H.
Ibarra O.
Inverardi P.
Jacquet Ph.
Jeffrey A.
Jones N. D.
Josko B.
Jourdan M.
Karhumäki J.
Kari J.
Kari L.
Karpinski M.
Katz S.
Kennaway R.
Kenyon C.
Kiehn A.
Kirschenhofer P.
Klarlund N.
Klasing R.
Klop J. W.
Klusener S.
Korach E.
Korec I.
Kortelainen J. M.
Kozen D. C.
Krasucki P.
Krob D.
Kubiak R.
Kuich W.
Kuiper R.
Künneke
Kwiatkowska M. Z.
Larsen K.
Latella D.
Latteux M.
Li A.
Lichtenberger F.
Lin H.
Linna M.
Litman A.
Litovsky I.
Louchard G.
Lüling R.
Maggiolo-Schettini A.
Marchetti-Spaccamela A.
Martelli A.
Masini A.
Mauri G.
Mayoh B. H.
Mazurkiewicz A.
Mehlhorn K.
Meir S.
Melichar B.
Menzel K.
Meyer auf der Heide F.
Middeldorp A.
Mignotte M.
Miltersen P. B.

Miola A.
Mnuk M.
Monien B.
Mücke E.
Mulder H.
Mumbeck W.
Murphy D. V. J.
Mysliwietz
Nanni U.
Naor J.
Nemes I.
Nielsen M.
Nielson F.
Nielson H.
Niemi V.
Niwiński D.
Olderog E.-R.
Orponen P.
Panangaden P.
Pantziou G.
Panzieri F.
Parrow J.
Paterson M. S.
Pawłowski W.
Paz A.
Peine R.
Peleg D.
Pelletier M.
Penaud J. G.
Penczek W.
Perrin D.
Pfalzgraf J.
Pighizzini G.
Pin J. E.
Pinkas B.
Plesnik J.
Pocchiola M.
Ponse A.
Preilowski
Privara I.
Prodinger H.
Protasi M.
Pultr A.
Quaglia P.
Ramme F.
Ramos E.
Régnier M.
Renvall A.
Reutenauer C.
Robert Ph.
Rolletschek H.
Roman J.
Römke Th.
Roscoe A. W.
Rose K.
Rosendahl M.
Rosenstiehl P.
Rössig S.
Roth R. M.
Rovan B.
Rutten J.
Ruzicka P.
Rydeheard D.
Sacibra A.
Sakarovitch J.
Salomaa A.
Salomaa K.
Sangiorgi D.
Santha M.
Schaffer R.
Schenke M.
Schmidt E. M.
Schöning U.
Schreiner W.
Schuster A.
Senizergues G.
Shachnai H.
Shah N.
Shapiro E.
Skyum S.
Slobodova A.
Snyder W.
Sokołowski St.
Sondergaard H.
Spirakis P.
Steffen B.
Steinby M.
Stern J.
Steyaert J. M.
Stifter S.
Stirling C.
Stokkermans K.
Stoughton A.
Sturc J.
Sykora O.
Szałas A.
Szpankowski W.
Talamo M.
Tan T. S.
Tarlecki A.
Teillaud M.
Tel G.
Tirze H.
Tofte M.
Tofts C.
Toyama Y.
Tran N.
Turakainen P.
Tzakalidis A.
Ukkonen E.
Unger W.
Van Draager F.
Van Oostrom V.
Vauquelin B.
Vollmer H.
Vrto I.
Wagner K.
Wagner K. W.
Walker D.
Wanke E.
Waupotitsch R.
Weil P.
Wernisch L.
Wiedermann J.
Williams P.
Winkler F.
Winkowski J.
Winskel G.
Zantema H.
Zielonka W.
Zwiers J.

CONTENTS

Complexity (Session 1)

Formal Languages (Session 2)

Finite Automata (Session 3)

Graph Grammars and Complexity (Session 4)

Algorithm Analysis (Session 5)

Algorithms (Session 6)

Parallel Computation (Session 7)

Graph Algorithms (Session 8)

Symbolic Computation (Session 9)

Geometric Algorithms (Session 10)

Logic and Models I (Session 11)

Logic and Models II (Session 12)

Time Specification (Session 13)

Concurrency Semantics (Session 14)

Program Development (Session 15)

Process Equivalences (Session 16)

Philosophical Issues in Kolmogorov Complexity

Ming Li[*] Paul M.B. Vitányi[†]
University of Waterloo 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 $(1 \le i \le j)$ 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

$$f(n) \le (2n/\log n)[1 + O(\log\log n/\log n)]. \qquad (1)$$

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

PROOF. Choose M such that $C(M|D_2, ..., D_j) \ge n$. 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 $1 \le i \le j$, m_i can be described using its discrepancy with $d_i/2$, hence

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

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

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

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, $C(\omega_{1:n}) \leq \log n + c$.