Electronic Colloquium on Computational Complexity, Report No. 30 (2004)

Kolmogorov complexity of enumerating finite sets

Nikolai K. Vereshchagin*

Keywords: Kolmogorov complexity, the a priori probability.

Abstract

Solovay [3] has proven that the minimal length of a program enumerating a set A is upper bounded by 3 times the absolute value of the logarithm of the probability that a random program will enumerate A. It is unknown whether one can replace the constant 3 by a smaller constant. In this paper, we show that the constant 3 can be replaced by the constant 2 for *finite* sets A.

We recall first two complexity measures ("information content") of computably enumerable sets defined by Solovay [3].

Let M be a machine with one infinite input tape and one infinite output tape. At the start the input tape contains an infinite binary string ω called the input to M. The output tape is empty at the start. We say that a program p *enumerates* a set $A \subset \mathbb{N}$ if in the run on every input ω extending p machine M prints all the elements of A in some order and no other elements. We do not require M to halt in the case when A is finite.[1] Let $KE_M(A)$ denote the minimal length of a program enumerating A. There is a machine M_0 (called a *universal* machine) such that for every other machine M there is a constant c such that

$$KE_{M_0}(A) \leq KE_M(A) + c$$

for all $A \subset \mathbb{N}$. Fix any such M_0 and call $KE(A) \stackrel{def}{=} KE_{M_0}(A)$ the *complexity of enumeration of* A. This complexity thus depends on the choice of the universal machine but this dependence is rather weak: for any other universal machine M_1 the difference $|KE_{M_0}(A) - KE_{M_1}(A)|$ is bounded by a constant not depending on A.

Similar to the a priori probability distribution on finite strings (or integer numbers) Solovay [3] has defined the a priori probability distribution on enumerable sets. The definition is as follows.

*Moscow Lomonossov University, Leninskie Gory 1, Moscow 119992, Email: ver@mccme.ru. Work was done while visiting Laboratoire d'Informatique Fondamentale, Université de Provence. Supported in part by the RFBR grants 02-01-22001, 03-01-00475, 358.20003.1

[1] In the case of finite sets any such program is called an *implicit description of* A, as opposed to *explicit description of* A when M is required to halt after having printed the last element of A.

ISSN 1433-8092

Let M be a machine with one infinite input tape and one infinite output tape as described above. For every $A \subset \mathbb{N}$ consider the probability

$$m^e_M(A) = \Pr[M \text{ on input } \omega \text{ enumerates } A].$$

One can easily show that if $m^e_M(A) > 0$ then A is enumerable.

The class of distributions of such form has a maximal one up to a multiplicative constant. In other words, there is a machine M_1 (called *optimal*) such that for every machine M there is a constant c such that

$$c \cdot m^e_{M_1}(A) \geq m^e_M(A)$$

for all $A \subset \mathbb{N}$. Fix any such M_1 and call $m^e(A) \stackrel{def}{=} m^e_{M_1}(A)$ the *a priori* probability of enumerating A. The a priori distribution thus depends on the choice of the optimal machine but this dependence is also weak: for any other optimal machine M_2 both ratios $m^e_{M_1}(A)/m^e_{M_2}(A)$ and $m^e_{M_2}(A)/m^e_{M_1}(A)$ are bounded by a constant not depending on A.

It is easy to see that $2^{-KE(A)} = O(m^e(A))$. In other words, $-\log m^e(A) \leq KE(A) + O(1)$ for all A. Solovay [3] has proven that conversely $KE(A) \leq -3\log m^e(A) + O(\log(-\log m^e(A))$ for all A. It is unknown whether we can replace the constant 3 in this inequality by a smaller constant. In this paper, we show that the constant 3 can be replaced by the constant 2 for *finite* sets A.

Theorem 1. *There is a constant c such that for every finite set A we have $KE(A) \leq -2\log m^e(A) + 2\log(-\log(m^e(A))) + c$.*

The proof is based on the ideas used to prove a lemma of Martin from [2]. The statement of Martin's lemma was used also in the Solovay's proof. In contrast, we are unable to use only the statement of the lemma.

Proof. Let $k = \lceil -\log m^e(A) \rceil + 1$. Given k we will enumerate $K = 2^k(2^k+1)/2$ sets $C_1, \ldots, C_K$ such that each finite set B with $m^e(B) \geq 2^{-k+1}$ coincides with some C_i. There is a machine M' that on every input beginning with

$$0^l 1(\text{binary notation of } k)(\text{binary notation of } i)$$

enumerates C_i, where l stands for the length of the binary notation of k. For this machine it holds $KE_{M'}(C_i) \leq 2l + 1 + \log K$ and by universality $KE(C_i) \leq KE_{M'}(C_i) + O(1) \leq \log K + 2l + O(1)$ for all i. As $m^e(A) \geq 2^{-k+1}$, we obtain

$$KE(A) \leq \log K + 2l + O(1) \leq 2k + 2\log k + O(1).$$

To enumerate $C_1, \ldots, C_K$ we run the optimal machine M defining m^e in steps and try all possible finite inputs to M. Say, on the stage t, we make t steps of the run of M on all inputs p of length t. Let $M^t(p)$ stand for the set enumerated by M in t steps on input p of length t (note that M cannot read in t steps more than t symbols from its input tape). Let Ω stand for the set of all infinite binary sequences and Ω_p for those beginning with the finite sequence p.

For each finite set B and on each stage t consider the set $S(B) = S^t(B) \subset \Omega$ that is the union of Ω_p over all p of length t such that $M^t(p) = B$. Note that $S(B)$ can both increase and decrease on stage t. Indeed, assume that $M^{t-1}(p) = B$ and on step t of the run on input p of length t the machine M writes a new element b on the output tape. Then $S(B)$ decreases on stage t: $S^t(B) = S^{t-1}(B) \setminus \Omega_p$, while $S^t(B \cup \{b\})$ increases: $S^t(B \cup \{b\}) = S^{t-1}(B \cup \{b\}) \cup \Omega_p$. Without loss of generality we may assume that on stage t this happens only for one pair (p, b). Otherwise we can split the stage into several substages.

Observing $S(B)$ for different B's we will enumerate sets $C_1, \dots, C_K \subset \mathbb{N}$. At each stage t we will enumerate a finite number of elements in some of $C_1, \dots, C_K$ so that at the end of stage t the following be true

$$\text{every finite set } B \text{ with } \lambda(S(B)) \geq 2^{-k} \text{ coincides with } C_i \text{ for some } i \leq K \quad (1)$$

where λ denotes the uniform measure on Ω.

Let us prove first that it suffices to keep true (1). Assume that B is a finite set such that $m^e(B) \geq 2^{-k+1}$. We claim that $m^e(B) = \lim_{t\to\infty} \lambda(S^t(B))$. Indeed, the set $S^t(B)$ is the difference of two sets: $S_1^t(B) = \{\omega \mid M(\omega)$ prints in at most t steps all the elements of $B\}$ and $S_2^t(B) = \{\omega \mid M_1(\omega)$ prints in at most t steps all the elements of B and an element outside $B\}$. Let $S_1^\infty(B)$ be the union of all $S_1^t(B)$ and $S_2^\infty(B)$ the union of all $S_2^t(B)$. As the uniform measure is continuous we have

$$\begin{aligned}
\lambda(S_1^\infty(B)) &= \lim_{t\to\infty} \lambda(S_1^t(B)) \\
\lambda(S_2^\infty(B)) &= \lim_{t\to\infty} \lambda(S_2^t(B))
\end{aligned}$$

and

$$\begin{aligned}
m^e(B) &= \lambda(S_1^\infty(B) \setminus S_2^\infty(B)) \\
&= \lambda(S_1^\infty(B)) - \lambda(S_2^\infty(B)) \\
&= \lim_{t\to\infty} \lambda(S_1^t(B)) - \lim_{t\to\infty} \lambda(S_2^t(B)) \\
&= \lim_{t\to\infty} (\lambda(S_1^t(B)) - \lambda(S_2^t(B))) \\
&= \lim_{t\to\infty} \lambda(S_1^t(B) \setminus S_2^t(B)) = \lim_{t\to\infty} \lambda(S^t(B)).
\end{aligned}$$

Therefore for almost all t we have $\lambda(S^t(B)) \geq 2^{-k}$. By (1) this implies that for almost all t there is i such that B coincides with C_i. Therefore there is i such that for infinitely many t we have $C_i = B$. Since C_i increases as t increases, this obviously implies that B coincides with C_i.

Now we need to explain how to enumerate $C_1, \dots, C_K$ to keep true condition (1). Let us call numbers in the segment $\{1, \dots, K\}$ *inspectors*. On each stage t, we assign to each inspector i its *rank*, a number in the segment $\{1, 2, \dots, K\}$. Also we assign to each inspector i a subset of Ω of the measure 2^{-k} called the *set controlled by* inspector i on stage t. At the end of each stage the ranks and controlled sets will satisfy the following invariant.

1. For all $r \leq 2^k$ there are exactly r different inspectors of rank r.
2. The sets controlled by different inspectors of the same rank are disjoint. As there are 2^k inspectors of rank 2^k, this item implies that the sets controlled by inspectors of rank 2^k form a partition of Ω.
3. If the set controlled by inspector i intersects with $S^t(B)$ then $C_i \subset B$.
4. For every finite B with $\lambda(S(B)) \geq 2^{-k}$ there is an inspectors i with $C_i = B$ (condition (1)).

We start with empty $C_1, \ldots, C_K$ and the ranks are assigned somehow to satisfy item 1. The controlled sets are also defined somehow so that item 2 be true. The items 3 and 4 are fulfilled, as all $C_1, \ldots, C_K$ are empty and $S(B)$ is non-empty only for $B = \emptyset$.

Let us proceed to the stage t. Assume that on stage t the set $S(B)$ decreases by Ω_p: $S^t(B) = S^{t-1}(B) \setminus \Omega_p$, while $S(B \cup \{b\})$ increases by Ω_p: $S^t(B \cup \{b\}) = S^{t-1}(B \cup \{b\}) \cup \Omega_p$. Recall that we assume that this happens only for one pair (p, b). (If this happens for no (p, b) we do nothing, as the invariant remains true in that case.)

As $S(B \cup \{b\})$ has increased, the item 4 may become false for the set $B \cup \{b\}$. Let us prove first that this is the only possible violation of the invariant. Item 1 remains true, since we have not yet changed the ranks. Item 2 remains true, since we have not yet changed the controlled sets. Let us prove that the item 3 remains true. Assume that the set controlled by inspector i intersects with $S^t(B')$. If B' is different from $B \cup \{b\}$ then it intersects also with $S^{t-1}(B') \supset S^t(B')$ and, since item 3 was true at the end of stage $t-1$ it remains true for B'. Assume that $B' = B \cup \{b\}$. As $S^t(B \cup \{b\}) \subset S^{t-1}(B) \cup S^{t-1}(B \cup \{b\})$, the set controlled by inspector i intersects with $S^{t-1}(B)$ or with $S^{t-1}(B \cup \{b\})$. As item 3 was true at the end of stage $t-1$ C_i is included in B or $B \cup \{b\}$. In both cases it is included in $B \cup \{b\}$.

Now we explain how to fulfill item 4 for $B \cup \{b\}$ in the case $\lambda(S(B \cup \{b\})) \geq 2^{-k}$. Choose any part T of $S(B \cup \{b\})$ of measure 2^{-k}. Let C_j be an inspector of the lowest rank r whose controlled set intersects with T (there is such an inspector, as the parts controlled by inspectors of rank 2^k form a partition of Ω). Decrease by 1 the rank of all inspectors of rank r except C_j and simultaneously increase by 1 the rank of all inspectors of rank $r-1$. Now the sets controlled by all inspectors of rank r except C_j are disjoint with T and we make C_j control T. So the item 2 remains true, as well as item 1. By item 3 the set C_j is included in $B \cup \{b\}$. Enumerate the difference $B \cup \{b\} \setminus C_j$ into C_j. The item 4 is now true for $B \cup \{b\}$. However, as C_j has been changed, item 4 may become false for B' equal to the previous content of C_j. The point is that this can happen only when B' is a proper subset of $B \cup \{b\}$. Apply the same procedure to B'. Again the item item 4 may become false only for one B'' that is a proper subset of B'. Hence after a finite number of applications of this procedure we restore item 4 for all sets. □

Acknowledgment. The author is sincerely grateful to Sergei Salnikov for writing down a preliminary version of the proof and for careful reading the paper.

References

[1] M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, New York, 2nd Edition, 1997.

[2] D.A. Martin, Borel indeterminacy, Ann. Math. 102 (1978) 363–371.

[3] R.M.Solovay, In: A.I. Arruda, N.C.A. da Costa, R. Chaqui (Eds.) On Random R.E. Sets, Non-Classical Logics, Model Theory and Computability, North-Holland, Amsterdam, 1977, pp. 283–307.

ECCC ISSN 1433-8092
http://www.eccc.uni-trier.de/eccc
ftp://ftp.eccc.uni-trier.de/pub/eccc
ftpmail@ftp.eccc.uni-trier.de, subject 'help eccc'