Kolmogorov complexity of enumerating finite sets

Nikolai K. Vereshchagin*

Keywords: Kolmogorov complexity, the a priori probability.

Abstract

Solovay [5] has proved 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 attributed by Solovay in [5] to G. Chaitin (wee keep Solovay's notations).

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 $A \subset \mathbb{N} = \{1, 2, \dots\}$ 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, and does not move the head on input tape beyond p. We do not require M to halt in the case when A is finite.[1] Let $I_M(A)$ 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

$$I_{M_0}(A) \le I_M(A) + c$$

for all $A \subset \mathbb{N}$. Fix any such M_0 and call $I(A) \overset{def}{=} I_{M_0}(A)$ 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 $|I_{M_0}(A) - I_{M_1}(A)|$ is bounded by a constant not depending on A.

The second complexity measure is related to the a priori probability distribution on enumerable sets. The definitions are as follows. Let M be a machine

*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 grant 06-01-00122-a

[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.

with one infinite input tape and one infinite output tape as described above. For every infinite 0-1-sequence ω let $M(\omega)$ denote the set enumerated by M when ω is written on its input tape. For every $A \subset \mathbb{N}$ consider the probability

$$m_M(A) = \Pr[M(\omega) = A].$$

A theorem of de Leeuw, Moore, Shannon and Shapiro [2] states that if $m_M(A) > 0$ 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

$$c \cdot m_{M_1}(A) \geq m_M(A)$$

for all $A \subset \mathbb{N}$. Fix any such M_1 and call $m(A) \stackrel{def}{=} m_{M_1}(A)$ 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 $m_{M_1}(A)/m_{M_2}(A)$ and $m_{M_2}(A)/m_{M_1}(A)$ are bounded by a constant not depending on A. Let $H(A)$ denote the negative binary logarithm of the a-priori probability of A: $H(A) = \lceil -\log m(A) \rceil$.

Comparing M_0, the machine defining $I(A)$, with M_1, the machine defining $m(A)$, we see that

$$H(A) = \lceil -\log m_{M_1}(A) \rceil \leq I_{M_1}(A) \leq I_{M_0}(A) + O(1) = I(A) + O(1)$$

for all A. Solovay [5] has proved that conversely $I(A) \leq 3H(A) + O(\log H(A))$ for all A, which can be viewed as a sharpening of de Leeuw et al.'s result.

Theorem 1 (Solovay). *There is a constant c such that for every set $A \subset \mathbb{N}$ we have $I(A) \leq 3H(A) + 2\log H(A) + c$.*

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 2. *There is a constant c such that for every finite set A we have $I(A) \leq 2H(A) + 2\log H(A) + c$.*

Our proof is basically a simplification of that of Solovay. Thus we first sketch Solovay's proof and then we present our proof.

Proof of Theorem 1 (a sketch). Fix k and let $A_1, \ldots, A_r$ be all the sets with $m(A) \geq 1/2^k$. Our goal is for every given k to enumerate $O(2^{3k})$ sets so that all $A_1, \ldots, A_r$ be among the enumerated sets.

First we need to understand the "constructive meaning" of the inequality $m(A) \geq 1/2^k$. Let Ω stand for the set of all infinite binary sequences and Ω_p for those beginning with the finite sequence p. Run the optimal machine M defining the a priori distribution m in steps and try all possible finite inputs to M. Say, we make t steps of the run of M on all inputs p of length t. (Note that

M cannot read in t steps more than t symbols from its input tape.) Let $M^t(\omega)$ stand for the set enumerated by M in t steps on input ω. For each $B \subset \mathbb{N}$ and t let $S^t(B) = \{\omega \mid M^t(\omega) = B\}$. The set $S^t(B)$ is *finitely based*, that is, it is a finite union of the sets having the form Ω_p. For all infinite B and all t the set $S^t(B)$ is empty. For finite B a code of $S^t(B)$ (a finite list of respective p's) can by computed given t and B. Note that $S^t(B)$ can both increase and decrease as t increases. Indeed, assume that $M^{t-1}(\omega) = B$ and on step t of the run on input ω of length t the machine M writes a new element b on the output tape. Let p be the prefix of ω of length t. Then $S(B)$ is decremented by Ω_p, while $S(B \cup \{b\})$ is incremented by Ω_p on step t.

Let μ denote the uniform measure on Ω and let $m^t(A) = \mu S^t(A)$. We can express m in terms of m^t as follows. For an integer l let $A[l]$ stand for the set of all $B \subset \mathbb{N}$ whose characteristic function coincides with that of A on l first naturals:

$$A[l] = \{B \subset \mathbb{N} \mid A \cap \{1, \dots, l\} = B \cap \{1, \dots, l\}\}.$$

The sequence of $A[l]$ decreases: $A[1] \supset A[2] \supset A[3] \supset \dots$ and its limit is equal to $\{A\} = \bigcap_l A[l]$. As m is a continuous measure, we obtain the following lemma, expressing $m(A)$ in terms of $m(A[l])$.

Lemma 1. *For all $A \subset \mathbb{N}$ we have $m(A) = \lim_{l\to\infty} m(A[l])$.*

The next lemma expresses $m(A[l])$ in terms of $m^t(A[l])$.

Lemma 2. *For all $A \subset \mathbb{N}$ and all l we have $m(A[l]) = \lim_{t\to\infty} m^t(A[l])$.*

Looking forward to prove Theorem 2 we state a lemma expressing m in terms of m^t for finite sets A more easily:

Lemma 3. *For all finite sets $A \subset \mathbb{N}$ we have $m(A) = \lim_{t\to\infty} m^t(A)$.*

It is because of Lemma 3 that we obtain a better bound than Solovay. Note that Lemma 3 is false for infinite sets A, as $m^t(A) = 0$ for all t in that case. We postpone the proofs of Lemmas 2 and 3 till the end of the paper, as well as the proofs of all other technical lemmas.

By Lemmas 1 and 2, if $m(A) \geq 2^{-k}$ then there are l, t with $m^t(A[l]) \geq 2^{-k-1}$. Say that a pair $\langle t, l\rangle$ is good for A if this happens. Say that a sequence of pairs $\langle t_1, l_1\rangle, \langle t_2, l_2\rangle, \dots$ is good for A if for infinitely many j the pair $\langle t_j, l_j\rangle$ is good for A. Solovay's algorithm can be split in a natural way into the following two algorithms.

Lemma 4. *There is an algorithm that given k and an auxiliary $k+1$-bit string λ_{k+1} computes a sequence of pairs $\langle t_1, l_1\rangle, \langle t_2, l_2\rangle, \dots$ that is good for all A with $m(A) \geq 2^{-k}$ and such that $l_1 < l_2 < l_3 < \dots$.*

Lemma 5. *There is an algorithm that given any sequence of pairs $\langle t_1, l_1\rangle, \langle t_2, l_2\rangle, \dots$ with $l_1 < l_2 < l_3 < \dots$ enumerates $N = O(2^{2k})$ sets $C_1, \dots, C_N$ such that every set A for which the sequence is good coincides with some C_i.*

Running the algorithm of Lemma 5 and using the algorithm of Lemma 4, as an external procedure, we obtain an algorithm that given k and λ_{k+1} enumerates sets $C_1, \dots, C_N$ satisfying Lemma 5. Thus there exists a machine that on every input ω beginning with

$$p = 0^{\log k}1(\text{binary notation of } k)(\lambda_{k+1})(\text{binary notation of } i)$$

scans p end then enumerates C_i. For this machine M' it holds

$$I_{M'}(C_i) \leq 2\log k + 1 + k + 1 + \log O(2^{2k})$$

and by universality

$$I(C_i) \leq I_{M'}(C_i) + O(1) \leq 3k + 2\log k + O(1)$$

for all i. Letting $k = H(A)$ and $C_i = A$ we obtain

$$I(A) \leq 3H(A) + 2\log H(A) + O(1).$$

Thus it remains to prove Lemmas 4 and 5.

Proof of Lemma 5. To construct the algorithm we need a computable strategy to win an infinite two person game defined in Martin's paper [4]. Let N, K be natural numbers. A configuration in this game is an N-tuple of finitely based subsets of Ω: $\langle X_1, \dots, X_N \rangle$. The initial configuration is $\langle \Omega, \dots, \Omega \rangle$. Player I on his turn plays a finitely based set Y with $\mu(Y) \geq 1/K$. If Y is disjoint with X_i for all $i = 1, \dots, N$, player I wins. If not, player II chooses a C_i intersecting Y and replaces C_i by Y, and the game continues. Player II wins if he can prevent I to win as described above for the entire, infinitely long game.

Martin [4] has proved that if $N = K(K+1)/2$ then player II has a computable winning strategy (uniformly in N, K).[2] We use Martin's result for $K = 2^{k+1}$.

Algorithm. We make steps $j = 1, 2, \dots$. At the end of step j we will have finite sets $C_1, \dots, C_N$ and a configuration $\langle X_1, \dots, X_N \rangle$ in Martin's game such that
(1) $C_i \subset \{1, \dots, l\}$ for all $i = 1, \dots, N$ (to simplify notation we omit subscript j in l_j, t_j);
(2) $\{B_1, \dots, B_s\} \subset \{C_1, \dots, C_N\}$, where $B_1, \dots, B_s$ are all subsets of $\{1, \dots, l\}$ for which $\langle t, l \rangle$ is good;
(3) $C_i \subset M^t(\omega)$ for all $\omega \in X_i$ and all $i = 1, \dots, N$.

At the start we let $C_i = \emptyset$ and $X_i = \Omega$ for all $i = 1, \dots, N$ thus conditions (1), (2) and (3) hold for $t = l = 0$.

Step j: find the pair $\langle t, l \rangle = \langle t_j, l_j \rangle$ and find all the sets $B_1, \dots, B_s$ such that $\langle t, j \rangle$ is good for B. Then play $Y = S^t(B_1[l])$ for player I in Martin's game. As $\mu S^t(B_1[l]) = m^t(B_1[l]) \geq 2^{-k-1}$ this is a legal move. Assume that the winning

[2] Later Ageev [1] showed that the condition $N = \Omega(K^2)$ is necessary for player I to win.

strategy of player II plays X_i. Let us show that $C_i \subset B_1$. As X_i intersects Y, there is $\omega \in X_i$ with

$$M^{t_j}(\omega) \cap \{1, \dots, l_j\} = B_1.$$

By condition (3) we have

$$C_i \subset M^{t_{j-1}}(\omega) \subset M^{t_j}(\omega).$$

By condition (1) this implies that

$$C_i \subset M^{t_j}(\omega) \cap \{1, \dots, l_j\} = B_1.$$

Update C_i by letting $C_i = B_1$. Note that now $C_i = B_1 \subset M^t(\omega)$ for all $\omega \in Y = X_i$ so condition (3) remains true for this i. Do the same for $B_2, \dots, B_s$ (note that $S^t(B_1[l]), \dots, S^t(B_s[l])$ are pairwise disjoint thus II's move always will be a new X_i). Note that the condition (3) remains true for all i such that C_i, X_i are unchanged, as $M^{t_j}(\omega) \supset M^{t_{j-1}}(\omega)$ for all ω. **End of Algorithm.**

Why does the Algorithm work? Assume that for infinitely many j the pair $\langle t_j, l_j \rangle$ is good for A. Then for infinitely many j there is i with $A \cap \{1, \dots, l_j\} = C_i$. As the number of C_i's is finite, we may assume that i is the same for infinitely many j. As l_j tends to infinity this implies that at the end we have $A = C_i$. □

Proof of Lemma 4. We first explain why we need extra $k+1$ bits of information. Try, say, $l_j = j$. By Lemma 1 for all large enough j for all A with $m(A) \geq 2^{-k}$ we have $m(A[l_j]) \geq 2^{-k-1} - 2^{-k-2}$. By Lemma 2 there is t_j with $m^{t_j}(A[l_j]) \geq 2^{-k-1}$ for all such A. The problem, however, is that the sequence of t_j might be uncomputable.

Roughly speaking, Solovay's idea to resolve this problem is as follows. Let $A_1, \dots, A_r$ be all sets with $m(A) \geq 2^{-k}$ and let $\lambda = \sum_{i=1}^{r} m(A_i)$. Let $B_1, \dots, B_s$ be all subsets of $\{1, \dots, l\}$ with $m^t(B[l]) \geq 2^{-k} - 1/n$ where t, l, n are large. If the sum $m^t(B_1[l]) + \dots + m^t(B_s[l])$ is close enough to λ then all the sets $A_i \cap \{1, \dots, l\}$ for $i = 1, \dots, r$ are in $\{B_1, \dots, B_s\}$, as otherwise the sum $\sum_{i=1}^{r} m(A_i)$ would be greater than λ.

More specifically, let λ_{k+1} denote the rational number consisting of $k+1$ first binary digits of λ. A triple $\langle t, n, l \rangle$ is called *opportune* if $n \geq 2^{k+1}$ and $m^t(B_1[l]) + \dots + m^t(B_s[l]) \geq \lambda_{k+1} - 2^{-k-1}$ where $B_1, \dots, B_s$ are all subsets of $\{1, \dots, l\}$ with $m^t(B[l]) \geq 2^{-k} - 1/n$.[3] The former condition implies that $m^t(B_i[l]) \geq 2^{-k-1}$. Since $B_1[l], \dots, B_s[l]$ are pairwise disjoint, we have $s \leq 2^{k+1}$.

Lemma 6. *For every c there is an opportune triple with $t, l, n > c$.*

Lemma 7. *If $m(A) \geq 2^{-k}$ and $\langle t_j, n_j, l_j \rangle$ is a sequence of opportune triples strictly increasing in each component then for infinitely many j we have*

$$m^{t_j}(A[l_j]) \geq 2^{-k} - 1/n_j \geq 2^{-k-1}.$$

[3]Solovay also requires that $m^t(B_1[l]) + \dots + m^t(B_s[l]) \leq \lambda_{k+1} + 2^{-k-1}$. This requirement is not necessary and omitting it simplifies the proof.

With these lemmas at hand we can easily finish the proof of Lemma 5. Fix k. Define the sequence $\langle t_j, n_j, l_j \rangle$ of opportune triples recursively: $\langle t_j, n_j, l_j \rangle$ is the first opportune triple that is strictly greater than the previous one in each component. Given k and λ_{k+1} we can compute $\langle t_j, n_j, l_j \rangle$ for every j. □

□

Proof of Theorem 2. Given k we enumerate $N = O(2^{2k})$ sets $C_1, \dots, C_N$ in steps $t = 1, 2, \dots$ so that at the end of step t every finite set B with $m^t(B) \geq 2^{-k-1}$ coincides with C_i for some $i \leq N$. We first establish that this will suffice for our purposes. If B is finite and $m(B) \geq 2^{-k}$ then by Lemma 3 for almost all t we have $m^t(B) \geq 2^{-k-1}$. 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 at the end.

Again we use Martin's game for $K = 2^{k+1}$. This time at the end of each step t the following two conditions will hold:
(1) $\{B_1, \dots, B_s\} \subset \{C_1, \dots, C_N\}$, where $B_1, \dots, B_s$ are all finite subsets of $\mathbb{N}$ with $m^t(B) \geq 2^{-k-1}$;
(2) $C_i \subset M^t(\omega)$ for all $\omega \in X_i$ and all $i = 1, \dots, N$.

At the start all C_i's are empty and the configuration in Martin's game is $\langle \Omega, \dots, \Omega \rangle$. On step t we find all $B_1, \dots, B_s$. Then we play $Y = S^t(B_1)$ for the player I in Martin's game. Let X_i be the move of the computable winning strategy of player II. As X_i intersects Y, there is $\omega \in X_i$ with $B_1 = M^t(\omega)$. By condition (2) we have $C_i \subset M^{t-1}(\omega) \subset M^t(\omega) = B_1$, hence $C_i \subset B_1$. We update C_i be letting $C_i = B_1$. Then we do the same for $B_2, \dots, B_s$. Note that $S^t(B_1), \dots, S^t(B_s)$ are pairwise disjoint. Therefore, the same C_i is not used twice. Now every B_i is among $C_1, \dots, C_N$. □

Proof of Lemma 2. The set $S^t(A[l])$ is the difference of two sets: $S_1^t = \{\omega \mid M(\omega)$ prints in at most t steps all the elements of $A \cap \{1, \dots, l\}\}$ and $S_2^t = \{\omega \mid M_1(\omega)$ prints in at most t steps all the elements of $A \cap \{1, \dots, l\}$ and an element in $\{1, \dots, l\} \setminus A\}$. Let S_1^∞ be the union of all S_1^t and S_2^∞ the union of all S_2^t. As the uniform measure is continuous we have

$$\mu(S_1^\infty) = \lim_{t \to \infty} \mu(S_1^t), \qquad \mu(S_2^\infty) = \lim_{t \to \infty} \mu(S_2^t),$$

and

$$\begin{aligned} m(A) &= \mu(S_1^\infty \setminus S_2^\infty) \\ &= \mu(S_1^\infty) - \mu(S_2^\infty) \\ &= \lim_{t \to \infty} \mu(S_1^t) - \lim_{t \to \infty} \mu(S_2^t) \\ &= \lim_{t \to \infty} (\mu(S_1^t) - \mu(S_2^t)) \\ &= \lim_{t \to \infty} \mu(S_1^t \setminus S_2^t) = \lim_{t \to \infty} \mu(S^t(A[l])). \quad \square \end{aligned}$$

Proof of Lemma 3. Let $S_1^t = \{\omega \mid M(\omega)$ prints in at most t steps all the elements of $A\}$ and $S_2^t = \{\omega \mid M_1(\omega)$ prints in at most t steps all the elements of A and an element outside $A\}$. The rest of the proof is the same as in the previous one. □

Proof of Lemma 6. Fix c. We need to find an opportune triple with $t, n, l > c$. Let n be any number greater than c and 2^{k+1}. Let l' be any number greater than c such that all sets $A_i \cap \{1, \dots, l'\}$ for $i = 1, \dots, r$ are pairwise different. By Lemma 1 we have $m(A_i) = \lim_l m(A_i[l])$ for all i. Let l be any number greater than l' such that $m(A_i[l]) > m(A_i) - 1/2nr$.

It suffices to show that if t is large enough then $m^t(A_i[l]) > 2^{-k} - 1/n$ for all $i = 1, \dots, r$ and $\sum_{i=1}^r m^t(A_i[l]) > \lambda_{k+1} - 2^{-k-1}$. By Lemma 2 for all large enough t we have $m^t(A_i[l]) > m(A_i[l]) - 1/2nr$ for all $i = 1, \dots, r$. Therefore $m^t(A_i[l]) > m(A_i[l]) - 1/2n \geq 2^{-k} - 1/n$ and $\sum_{i=1}^r m^t(A_i[l]) > \lambda - 1/2n \geq \lambda_{k+1} - 2^{-k-1}$. □

Proof of Lemma 7. Passing to a subsequence of $\langle t_j, n_j, l_j \rangle$ we may assume that the number s of different sets $B \subset \{1, \dots, l_j\}$ with $m^{t_j}(B[l_j]) \geq 2^{-k} - 1/n_j$ does not depend on j. Let B_i^j denote the value of B_i on step j. Again passing to a subsequence we may assume that $B_i^j \cap \{1, \dots, l_{j-1}\} = B_i^{j-1}$ for all $i = 1, \dots, s$ and all j. Let $B_i = \bigcup_j B_i^j$.

It suffices to prove that A coincides with some of $B_1, \dots, B_s$. To this end we show that $m(B_i) \geq 2^{-k}$ for all $i = 1, \dots, s$ and $m(\{B_1, \dots, B_s\}) \geq \lambda_{k+1} - 2^{-k-1}$ (notice that we do not assume that all $B_1, \dots, B_s$ are pairwise distinct). This implies that $A \in \{B_1, \dots, B_s\}$, as otherwise the sum of $m(B)$ over all distinct B with $m(B) \geq 2^{-k}$ would be greater that λ.

Let us prove first that $m(B_i) \geq 2^{-k}$. By way of contradiction assume that $m(B_i) < 2^{-k}$ and choose n' with $m(B_i) < 2^{-k} - 1/n'$. By Lemma 1 there is l' with $m(B_i[l']) < 2^{-k} - 1/n'$. By Lemma 2 there is t' such that $m^t(B_i[l']) < 2^{-k} - 1/n'$ for all $t > t'$. If $l_j > l'$, $t_j > t'$ and $n_j > n'$ then we have $m^{t_j}(B_i^j[l_j]) = m^{t_j}(B_i[l_j]) \leq m^{t_j}(B_i[l']) < 2^{-k} - 1/n' < 2^{-k} - 1/n_j$, a contradiction.

It remains to prove that $m(\{B_1, \dots, B_s\}) \geq \lambda_{k+1} - 2^{-k-1}$. By Lemma 1 we have

$$m(\{B_1, \dots, B_s\}) = \lim_j m(B_1[l_j] \cup \dots \cup B_s[l_j]) = \lim_j m(B_1^j[l_j] \cup \dots \cup B_s^j[l_j]).$$

As the sets $B_1^j, \dots, B_s^j$ are distinct and $m(B_i^j[l_j]) = \lim_t m^t(B_i^j[l_j])$ we have

$$m(B_1^j[l_j] \cup \dots \cup B_s^j[l_j]) = \sum_i \lim_t m^t(B_i^j[l_j]) = \lim_t \sum_i m^t(B_i^j[l_j]).$$

Thus it suffices to prove that for all j,

$$\lim_t \sum_i m^t(B_i^j[l_j]) \geq \lambda_{k+1} - 2^{-k-1}.$$

Fix j. To prove the inequality it is enough to present an infinite sequence of t's with

$$\sum_i m^t(B_i^j[l_j]) \geq \lambda_{k+1} - 2^{-k-1}.$$

By construction the sequence $t_j, t_{j+1}, \ldots$ has this property, as

$$\sum_i m^{t_{j+a}}(B_i^j[l_j]) \geq \sum_i m^{t_{j+a}}(B_i^{j+a}[l_{j+a}]) \geq \lambda_{k+1} - 2^{-k-1}. \quad \square$$

Acknowledgments. The author is sincerely grateful to Sergei Salnikov for writing down a preliminary version of the proof and for the anonymous referees for helpful suggestions.

References

[1] M. Ageev, Martin's game: a lower bound for the number of sets, Theor. Comput. Sci. 289(1): 871-876 (2002)

[2] K. de Leeuw, E. F. Moore, C. E. Shannon, and N. Shapiro, Computability by probabilistic machines, In: C. E. Shannon and J. McCarthey (Eds.), Automata Studies, Princeton University Press, Princeton, New Jersey, 1956, 183–212.

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

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

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