CDMTCS Research Report Series

Minimal Programs Are Almost Optimal

Cristian S. Calude
Auckland University, New Zealand

Hajime Ishihara
JAIST, Japan

Takeshi Yamaguchi
Tokyo Denki University, Japan

CDMTCS-116
November 1999

Centre for Discrete Mathematics and
Theoretical Computer Science

Minimal Programs Are Almost Optimal*

Cristian S. Calude† Hajime Ishihara‡ Takeshi Yamaguchi§

Abstract–**According to the Algorithmic Coding Theorem, minimal programs of any universal machine are prefix-codes asymptotically optimal with respect to the machine algorithmic probabilities. A stronger version of this result will be proven for a class of machines, not necessarily universal, and any semi-distribution. Furthermore, minimal programs with respect to universal machines will be shown to be almost optimal for any semi-computable semi-distribution. Finally, a complete characterization of all machines satisfying the Algorithmic Coding Theorem is given.**

Indexed Terms–**Minimal program, prefix-code, Chaitin machine, program-size complexity, entropy, Noiseless Coding Theorem, Algorithmic Coding Theorem.**

1 Introduction

Let C be a prefix-code with one code string per source string, that is, an one-one function from binary strings to binary strings whose range is a prefix-free set. Let $P(x)$ be the probability of the source string x and let $|C(x)|$ be the length of the code string of x. Shannon's Noiseless Coding Theorem says that the minimal average code string length is about equal to the entropy of the source string set. The strategy is to chose a prefix-code that matches best the probability distribution of the source codes.

In what follows we will be interested in infinite prefix-codes, that is, prefix-codes naming all binary strings. We will work with semi-distributions, i.e., functions P from strings to reals such that $\sum_x P(x) \leq 1$. A Shannon type result is valid for semi-distributions. We will be interested in finding prefix-free codes which are almost optimal for a given semi-distribution, and also for a class of semi-distributions (in case the semi-distribution may be unknown, or uncomputable).

Algorithmic Information Theory (see [3, 4, 1, 5]) provides a natural class of prefix-free codes, namely the set of minimal (canonical) programs of a machine. These prefix-codes are important for statistical physics, [8]. According to Algorithmic Coding Theorem, minimal programs of any universal machine (a machine capable of simulating any other machine) are asymptotically optimal with respect to the machine algorithmic probabilities. A stronger version of this result is proven for a class of machines, not necessarily universal, and any semi-distribution. Minimal programs with respect to universal machines are proven almost optimal for any semi-computable semi-distribution. Finally, a complete characterization of all machines satisfying the Algorithmic Coding Theorem is given.

---

*Calude's work was done during his visit to Jaist as a Monbusho visiting professor. Ishihara was partly supported by a Grant-in-Aid for Scientific Research (C) No 09640253 of the Japan Ministry of Education, Science, Sports and Culture.

†School of Information Science, Japan Advanced Institute of Science, and Technology, JAIST, Tatsunokuchi, Ishikawa 923-1292, Japan; on leave from the Department of Computer Science, University of Auckland, Private Bag 92019, Auckland, New Zealand; e-mails: cristian@{jaist.ac.jp,cs.auckland.ac.nz}.

‡School of Information Science, Japan Advanced Institute of Science, and Technology, JAIST, Tatsunokuchi, Ishikawa 923-1292, Japan; e-mail: ishihara@jaist.ac.jp.

§Department of Information Sciences, College of Science and Engineering, Tokyo Denki University, Hatoyama-Machi, Hiki-Gun, Saitama-Ken 350-0394, Japan; e-mail: y-takesi@j.dendai.ac.jp.

2 Notation, Definitions and Basic Results

By $\mathbf{N}, \mathbf{Q}$, and Σ^* we denote the sets of non-negative integers, rationals, and (finite) binary strings, respectively. The length of a string s is denoted by $|s|$. A string s is a prefix of a string t ($s \subseteq t$) if t is the concatenation of $s \in \Sigma^*$ and some string r. A subset A of Σ^* is *prefix-free* if whenever s and t are in A and $s \subseteq t$, then $s = t$. For example, the set $\{1^i 0 \mid i \geq 0\}$ is prefix-free. *Kraft's inequality* states that for every prefix-free set $A \subset \Sigma^*$, $\sum_{s \in A} 2^{-|s|} \leq 1$.

By log we denote the base 2 logarithm. For every real α, if $2^n < \alpha \leq 2^{n+1}$, for some integer n, then put $n = \lg \alpha$. Note that if $\alpha > 0$, then $2^{\lg \alpha} < \alpha, \lg \alpha < \log \alpha \leq \lg \alpha + 1$, and if m is an integer, then $\lg \alpha \geq m$ iff $\log \alpha > m$.

We assume familiarity with Turing machines, computable sets and functions, computably enumerable (c.e.) sets, e.g., from [10]. We shall employ a sspecial model of deterministic Turing machine computation, namely *self-delimiting Turing machines* or *(Chaitin) machines*: these are Turing machines (transforming binary strings into binary strings) and having prefix-free domains. More precisely, for every Chaitin machine M the program set $PROG_M = \{x \in \Sigma^* \mid M(x) \text{ halts}\}$ is a prefix-free. Note that, conversely, every prefix-free c.e. set set of strings is the domain of some Chaitin machine. In what follows we will operate only with Chaitin machines, which will be simply referred as machines.

The following result will be frequently used (see [2] for a simple proof):

Kraft-Chaitin Theorem. *Given a computable list of "requirements" $(n_i, s_i)(i \geq 0, s_i \in \Sigma^*, n_i \in \mathbf{N})$ such that $\sum_i 2^{-n_i} \leq 1$, we can effectively construct a machine M and a computable one-to-one enumeration $x_0, x_1, x_2, \ldots$ of strings x_i of length n_i such that $M(x_i) = s_i$ for all i, and $M(x)$ is undefined if $x \notin \{x_i \mid i \in \mathbf{N}\}$.*

The *program-size complexity* induced by the machine M is

$$H_M(x) = \min\{|z| \mid M(z) = x\},$$

with the convention that the minimum of the empty set is undefined.

The *algorithmic probability* of the machine M to produce the output x is

$$P_M(x) = \sum_{M(u)=x} 2^{-|u|},$$

and the halting probability of M is

$$\Omega_M = \sum_{x \in \Sigma^*} P_M(x) = \sum_{x \in PROG_M} 2^{-|x|}.$$

It follows by Kraft's inequality that, for every machine M and any string $x \in \Sigma^*$,

$$0 \leq P_M(x) \leq \Omega_M \leq 1.$$

For every machine M and string x such that $P_M(x) > 0$, we denote by

$$x^*_M = \min\{u \mid M(u) = x\},$$

where the minimum is taken according to the quasi-lexicographical ordering of strings; x^*_M is called the *minimal (canonical) program* of x with respect to M.

A machine U is *universal* if for every machine M, there is a constant c_M (depending upon M) with the following property: if $M(x)$ halts, then there is an $x' \in \Sigma^*$ such that $U(x') = M(x)$ and $|x'| \leq |x| + c_M$; c_M is the simulation constant of M on U. Universal machines can be effectively constructed. If U is universal, then x^*_U exists for every string x. See more in [1].

Algorithmic Coding Theorem. (Chaitin) *There exists a constant $c \geq 0$ such that for all strings x,*

$$|H_U(x) + \log P_U(x)| \leq 1 + c.$$

See [3, 4, 1, 5].

3 Noiseless Coding Theorem

A function $P : \Sigma^* \to [0,1]$ such that $\sum_x P(x) \leq 1$ is called a *semi-distribution* over the strings. In case $\sum_x P(x) = 1$, P is a *distribution*. A semi-distribution P is semi-computable from below (above) in case the set $\{(x,r) \mid x \in \Sigma^*,\ r \in \mathbf{Q},\ P(x) > r\}$ ($\{(x,r) \mid x \in \Sigma^*,\ r \in \mathbf{Q},\ P(x) < r\}$) is c.e. A semi-distribution P is computable if it is semi-computable from below and from above. For example, P_M is a semi-distribution semi-computable from below. The function $P(x) = 2^{-2|x|-1}$ is a computable distribution.

A *code* for strings is an one-one function $C : \Sigma^* \to \Sigma^*$ such that $C(\Sigma^*)$ is prefix-free. For example, for every surjective machine M, $C_M(x) = x^*_M$ is a code; universal machines are surjective. The *average code-string length* of a code C with respect to a semi-distribution P is

$$L_{C,P} = \sum_x P(x) \cdot |C(x)|.$$

The *minimal average code-string length* with respect to a semi-distribution P is

$$L_P = \min\{L_{C,P} \mid C \text{ code}\}.$$

The *entropy* of a semi-distribution P is

$$\mathcal{H}_P = -\sum_x P(x) \cdot \log P(x).$$

A simple extension of Shannon's classical argument [9] (see more in [7]) leads to the following

Noiseless Coding Theorem. (Shannon) *The following inequalities hold true for every semi-distribution P:*

$$\begin{aligned} \mathcal{H}_P - 1 &\leq \mathcal{H}_P + \left(\sum_x P(x)\right) \log \left(\sum_x P(x)\right) \\ &\leq L_P \leq \mathcal{H}_P + 1. \end{aligned}$$

If P is a distribution, then $\log(\sum_x P(x)) = 0$, so we get the classical inequality $\mathcal{H}_P \geq L_P$. However, this inequality is not true for every semi-distribution. For example, take $P(x) = 2^{-2|x|-3}$ and $C(x) = x_1 x_1 \ldots x_n x_n 01$. It follows that $L_P \leq L_{C,P} = \mathcal{H}_P - \frac{1}{4}$.

4 Main Result

We investigate conditions under which given a semi-distribution P, we can find a (universal) machine M such that $H_M(x)$ is equal, up to an additive constant, to $-\log P(x)$. In what follows we will assume that $P(x) > 0$, for every x.

We start with the main technical result:

Theorem 1 *Assume that P is a semi-distribution and there exist a c.e. set $S \subset \Sigma^* \times \mathbf{N}$ and a constant $c \geq 0$ such that the following two conditions are satisfied for every $x \in \Sigma^*$:*

1) $\sum_{(x,n) \in S} 2^{-n} \leq P(x)$,
2) *if $P(x) > 2^{-n}$, then $(x,m) \in S$, for some $m \leq n + c$.*

Then, there exists a machine M (depending upon S) such that for all x,

$$-\log P(x) \leq H_M(x) \leq (1+c) - \log P(x). \quad (1)$$

Proof. In view of 1),

$$\sum_{(x,n) \in S} 2^{-n} \leq \sum_x P(x) \leq 1,$$

so using Kraft-Chaitin Theorem we can construct a machine M such that for every $(x,n) \in S$ there exists a string $v_{x,n}$ of length n such that $M(v_{x,n}) = x$. Using 1) and 2), we get:

$$\begin{aligned} H_M(x) &= \min\{|v| \mid v \in \Sigma^*, M(v) = x\} \\ &= \min\{n \mid n \in \mathbf{N}, (x,n) \in S\} \\ &\leq \min\{m \mid m \in \mathbf{N}, P(x) > 2^{-m}\} + c \\ &= \min\{m \mid m \in \mathbf{N}, m > -\log P(x)\} + c \\ &= \min\{m \mid m \in \mathbf{N}, m \geq 1 - \lg P(x)\} + c \\ &\leq (1+c) - \log P(x). \end{aligned}$$

□

Remark. Theorem 1 makes no computability assumptions on P.

Lemma 2 *Let M be a machine such that $\Omega_M < 1$. Then, there exists a universal machine U satisfying the inequality $H_U(x) \leq H_M(x)$, for all x.*

Proof. By hypothesis, $\Omega_M < 1$, so there is a non-negative integer k such that $\Omega_M + 2^{-k} \leq 1$. Let V be a universal machine. The set

$$\begin{aligned} S &= \{(M(x), |x|) \mid x \in PROG_M\} \\ &\quad \cup \{(V(x), |x| + k) \mid x \in PROG_V\} \end{aligned}$$

is c.e. and

$$\sum_{(y,n) \in S} 2^{-n} \leq \Omega_M + 2^{-k} \leq 1.$$

Consequently, in view of Kraft-Chaitin Theorem, there exists a machine U such that for $(y, n) \in S$ there is a program $z \in PROG_U$ of length n such that $U(z) = y$. Clearly, for every x,

$$H_U(x) \leq \min\{|w| + k \mid V(w) = x\} = H_V(x) + k,$$

and

$$H_U(x) = \min\{|v| \mid U(v) = x\} \leq H_M(x),$$

so U is universal and satisfies the required inequality. □

Lemma 3 *Let M be a machine. Then, there exists a machine M' such that $\Omega_{M'} < 1$ and $H_{M'}(x) = H_M(x) + 1$, for all x.*

Proof. Apply Kraft-Chaitin Theorem to the set $\{(M(x), |x| + 1) \mid x \in PROG_M\}$ to obtain the machine M'. □

Corollary 4 *Under the hypotheses of Theorem 1, a universal machine U can be constructed such that for all x,*

$$H_U(x) \leq (2 + c) - \log P(x). \qquad (2)$$

Proof. Use Lemmas 3, 2 to get a universal machine U such that $H_U(x) \leq H_M(x) + 1$, for all x. □

5 Coding with Minimal Programs

Specializing P in Theorem 1 we show that minimal programs are almost optimal for P. Minimal programs of universal machines are almost optimal for every semi-computable semi-distribution P.

Semi-computable semi-distributions from below (e.g., algorithmic probabilities of machines) are important in Algorithmic Information Theory.

Proposition 5 *Assume that P is a semi-distribution semi-computable from below. Then, there exists a machine M (depending upon P) such that for all x,*

$$-\log P(x) \leq H_M(x) \leq 2 - \log P(x). \qquad (3)$$

Consequently, minimal programs for M are almost optimal: the code C_M satisfies the inequalities:

$$0 \leq L_{C_M,P} - \mathcal{H}_P \leq 2.$$

Proof. Take $S = \{(x, n + 1) \mid P(x) > 2^{-n}\}$. For every x we have:

$$\sum_{(x,n) \in S} 2^{-n} = \sum_{n > 1 - \log P(x)} 2^{-n} = \sum_{n \geq 1 - \lg P(x)} 2^{-n}$$

$$= 2^{\lg P(x)} < P(x),$$

so condition 1) in Theorem 1 is satisfied. Condition 2) holds for $c = 1$. Finally,

$$L_{C_M,P} - \mathcal{H}_P = \sum_x P(x) \cdot (H_M(x) + \log P(x)) \leq 2.$$

□

Corollary 6 *Assume that $f : \Sigma^* \to \mathbf{N}$ is a function such that the set $\{(x, n) \mid f(x) < n\}$ is c.e. and $\sum_x 2^{-f(x)} \leq 1$. Let $P(x) = 2^{-f(x)}$. Then P is a semi-distribution semi-computable from below, and there exists a machine M (depending upon f) such that for all x,*

$$H_M(x) \leq 1 + f(x). \qquad (4)$$

Minimal programs for M are almost optimal: the code C_M satisfies the inequalities:

$$0 \leq L_{C_M,P} - \mathcal{H}_P \leq 1.$$

One more bit is enough to guarantee universality of the constructed machine, that is, there exists a universal machine U (depending upon f) such that the code C_U satisfies the inequalities:

$$0 \leq L_{C_U,P} - \mathcal{H}_P \leq 2.$$

Proof. Take $S = \{(x,n) \mid n > f(x)\}$. Clearly, $S = \{(x,n) \mid P(x) > 2^{-n}\}$. The first condition in Theorem 1 is satisfied as $\sum_{n>f(x)} 2^{-n} = P(x)$, for every x, and the second condition is satisfied for $c = 0$. □

When the semi-distribution P is given, an optimal prefix-code can be found for P. However, that code may be far from optimal for a different semi-distribution. For example, let C pe a prefix-code such that $|C(x)| = 2^{|x|+2}$, for all x. Let $\alpha > 0$ and consider the distribution

$$P_\alpha(x) = (1 - 2^{-\alpha})\, 2^{-(\alpha+1)|x|}.$$

Two radically different situation appear: if $\alpha \leq 1$, then

$$L_{C,P_\alpha} - \mathcal{H}_{P_\alpha} = \infty,$$

but if $\alpha > 1$, then

$$L_{C,P_\alpha} - \mathcal{H}_{P_\alpha} < \infty.$$

So, C is asymptotical optimal for every distribution P_α with $1 < \alpha$, but C is far away from optimality if $0 < \alpha \leq 1$. Note that P_α is computable provided α is computable.

The next result shows that minimal programs are asymptotical optimal for every semi-distribution semi-computable from below.

Theorem 7 *Let P be a semi-distribution semi-computable from below, and U a universal machine. Then, there exists a constant c_P (depending upon P) such that*

$$0 \leq L_{C_U,P} - \mathcal{H}_P \leq 1 + c_P.$$

Proof. Take M the machine constructed in Proposition 5 and let c_M be the simulation constant of M on U. Then,

$$0 \leq L_{C_U,P} - \mathcal{H}_P \leq L_{C_M,P} + c_M - \mathcal{H}_P \leq 1 + c_M,$$

so take $c_P = c_M$. □

Remark. Corollary 7 generalizes a result in [6] proven for computable distributions. See also [8]. The result is important only for semi-distributions for which the entropy is infinite. For example, the entropy of the semi-distribution $P(x) = 2^{-|x|}\frac{1}{(|x|+2)\log(|x|+2)}$ is infinite.

Using Lemma 2 we can obtain sharper inequalities. For example, for every universal machine U, the code C_U is almost optimal with respect to P_U:

$$0 \leq L_{C_U,P_U} - \mathcal{H}_{P_U} \leq 2.$$

If f is a function as in Corollary 6 such that $\sum_x 2^{-f(x)} < 1$, then there exists a universal machine U such that

$$0 \leq L_{C_U,P} - \mathcal{H}_P \leq 1.$$

For example, take $f(x) = H_U(x)$, where U is a universal machine.

Proposition 8 *Let P be a computable semi-distribution. Then, there exists a machine M such that*

$$-\log P(x) \leq H_M(x) \leq 1 - \log P(x).$$

Proof. Note that $-\lg P(x) = \min\{n \mid n \in \mathbf{N},\ P(x) > 2^{-n}\}$ and then apply Theorem 1 to the set $S = \{(x, -\lg P(x)) \mid x \in \Sigma^*\}$ and constant $c = 0$. □

Corollary 9 *Let P be a computable semi-distribution. Then, there exists a universal machine U such that*

$$H_U(x) \leq 1 - \log P(x).$$

6 Algorithmic Coding Theorem Revisited

We characterize all machines satisfying the Algorithmic Coding Theorem and we construct a class of (universal) machines for which the inequality is satisfied with constant $c = 0$. This addresses the relevance of the theorem for statistical physics where the presence of an arbitrary constant is unsatisfactory (see [8]).

Proposition 10 *Let M be a machine and $c \geq 0$. The following statements are equivalent:*

1) *for all x, $H_M(x) \leq (1+c) - \log P_M(x)$,*

2) *for all non-negative n, if $P_M(x) > 2^{-n}$, then $H_M(x) \leq n + c$.*

Proof. From $H_M(x) \leq (1+c) - \log P_M(x)$ and $P_M(x) > 2^{-n}$ we deduce

$$2^{-n} < P_M(x) \leq 2^{(1+c)-H_M(x)}.$$

Conversely, we have:

$$H_M(x) - c = \min\{n \mid n \in \mathbf{N},\ P_M(x) > 2^{-n}\}.$$

□

For any machine M satisfying condition 2) in Proposition 10 the Algorithmic Coding Theorem holds:

$$|H_M(x) + \log P_M(x)| \leq 1 + c. \quad (5)$$

In fact, a machine M satisfies (5) if and only if condition 2) is satisfied. Every universal machine U satisfies condition 2), but not all machines satisfy this condition. To construct such an example, consider the following enumeration: for every string x enumerate $2^{|x|}$ copies of the pair $(x, 3|x|+1)$. Use Kraft-Chaitin Theorem to construct a machine M such that for every string x there exist $2^{|x|}$ different strings u_x^i, all of length $3|x|+1$, such that

$$M(u_x^i) = x,\ i = 1, 2, \ldots, 2^{|x|}.$$

It is seen that $P_M(x) = 2^{-2|x|-1}$, so taking $n_x = 2|x| + 2$ we get $P_M(x) > 2^{-n_x}$, but there is no constant c such that $H_M(x) \leq n_x + c$, for all strings x.

Some machines satisfy condition 2) with $c = 0$, so their canonical programs are almost optimal. A class of (universal) such machines is provided in the next proposition.

Proposition 11 *Let M be a machine such that for all programs $x \neq x'$ with $M(x) = M(x')$ we have $|x| \neq |x'|$. Then, for all x,*

$$H_M(x) \leq 1 - \log P_M(x). \quad (6)$$

Proof. Consider the set $S = \{(x, |y|) \mid M(y) = x\}$, and notice that

$$P_M(x) = \sum_{(x,n) \in S} 2^{-n},$$

as programs producing the same output have different lengths. In view of the hypothesis,

$$P_U(x) > 2^{-n} \iff \exists\,(x, k_1) \in S,\ k_1 < n \text{ or } (k_1 = n\ \&\ \exists\, k_2 \neq k_1,\ (x, k_2) \in S),$$

hence the second condition in Theorem 1 is satisfied with $c = 0$. Using Theorem 1 we deduce the existence of a machine M' such that $H_{M'}(x) \leq 1 - \log P_{M'}(x)$, for all x. The inequality (6) follows from

$$H_M(x) = \min\{n \mid (x, n) \in S\} = H_{M'}(x).$$

□

Remark. Not every universal machine satisfies the hypothesis of Proposition 11. However, if V is a universal machine, the one can effectively construct a universal machine U such that programs producing the same output via U have different lengths and $H_U(x) = H_V(x)$, for every x.[1] Indeed, enumerate the graph of V and as soon as a pair $(x, V(x))$ appears in the list do not include in the list any pair $(x', V(x'))$ with $x \neq x'$ and $V(x) = V(x')$. The set enumerated in this way, which is a subset of the graph of V, is the graph of the universal machine U satisfying the required condition.

[1] Of course, $P_U(x) \leq P_V(x)$, for all x.

References

[1] C. Calude. *Information and Randomness. An Algorithmic Perspective.* Springer-Verlag, Berlin, 1994.

[2] C. Calude, C. Grozea. Kraft-Chaitin inequality revisited, *J. Univ. Comput. Sci.* 5 (1996), 306–310.

number

[3] G. J. Chaitin. *Information, Randomness and Incompleteness, Papers on Algorithmic Information Theory*, World Scientific, Singapore, 1987. (2nd ed., 1990)

[4] G. J. Chaitin. *Algorithmic Information Theory*, Cambridge University Press, Cambridge, 1987. (third printing 1990)

[5] G. J. Chaitin. *The Unknowable*, Springer-Verlag, Singapore, 1999.

[6] T. M. Cover, P. Gács, R. M. Gray. Kolmogorov's contributions to information theory and algorithmic complexity, *The Annals of Probability* 17(1989), 840–865.

[7] T. M. Cover, J. A. Thomas. *Elements of Information Theory*, John Wiley, New York, 1991.

[8] R. Schack. Algorithmic information and simplicity in statistical physics, *Int. J. Theor. Physics*, 36 (1997), 209–226.

[9] C. E. Shannon. A mathematical theory of communication, *Bell. Sys. Tech. Journal* 27 (1948), 379–423, 623–656.

[10] R. I. Soare. *Recursively Enumerable Sets and Degrees*, Springer-Verlag, Berlin, 1987.