Universal Recursively Enumerable Sets of Strings

Cristian S. Calude[1], André Nies[2], Ludwig Staiger[3] and Frank Stephan[4⋆]

[1] Department of Computer Science, The University of Auckland, Private Bag 92019, Auckland, New Zealand, cristian@cs.auckland.ac.nz.

[2] Department of Computer Science, The University of Auckland, Private Bag 92019, Auckland, New Zealand, andre@cs.auckland.ac.nz.

[3] Martin-Luther-Universität Halle-Wittenberg, Institut für Informatik, D - 06099 Halle, Germany, staiger@informatik.uni-halle.de.

[4] Department of Mathematics and School of Computing, National University of Singapore, Singapore 117543, fstephan@comp.nus.edu.sg.

Abstract. The present work clarifies the relation between domains of universal machines and r.e. prefix-free supersets of such sets. One such characterisation can be obtained in terms of the spectrum function $s_W(n)$ mapping n to the number of all strings of length n in the set W. An r.e. prefix-free set W is the superset of the domain of a universal machine iff there are two constants c, d such that $s_W(n) + s_W(n+1) + \ldots + s_W(n+c)$ is between $2^{n-H(n)-d}$ and $2^{n-H(n)+d}$ for all n; W is the domain of a universal machine iff there is a constant c such that for each n the pair of n and $s_W(n) + s_W(n+1) + \ldots + s_W(n+c)$ has at least the prefix-free description complexity n. There exists a prefix-free r.e. superset W of a domain of a universal machine which is the not a domain of a universal machine; still, the halting probability Ω_W of such a set W is Martin-Löf random. Furthermore, it is investigated to which extend this results can be transferred to plain universal machines.

1 Introduction

The present paper provides a classification of recursively enumerable prefix codes using algorithmic information theory [1, 4–6, 10, 11]. The paper combines recursion theoretic arguments with (combinatorial) information theory. It is well-known that recursion theory does not yield a sufficiently fine distinction between several classes of recursively enumerable prefix codes, as, for example, the prefix code $S = \{0^n 1 : n \in W\}$ has the same complexity as the subset $W \subseteq N$ and all these prefix codes are indistinguishable by their entropy.

On the other hand one may assume that recursively enumerable prefix codes are in some sense "maximally complex" if they are the domains of universal prefix-free Turing machines. This observation is supported by Corollary 2 of [3] which states that every recursively enumerable prefix code is one-to-one embeddable into the domain of a universal prefix-free Turing machine by a partial recursive mapping increasing the output length at most by a constant. Moreover, this

⋆ F. Stephan is supported in part by NUS grant number R252-000-308-112. L. Staiger and F. Stephan are external researchers of the University of Auckland. A. Nies is partially supported by the Marsden Fund of New Zealand, grant no. 03-UOA-130.

characterisation yields a connection to the information-theoretic density of a universal recursively enumerable prefix code. Calude and Staiger [3] showed that universal recursively enumerable prefix codes have maximal density — in contrast with the code S discussed above.

The present paper provides a more detailed characterisation of the domains of universal prefix-free Turing machines and universal recursively enumerable (r.e.) prefix codes in terms of the spectrum function. More technically, the following definitions are made.

1. X is the set $\{0,1\}$ and X^* is the set of all strings over X: $X^* = \{\lambda, 0, 1, 00, 01, 10, 11, 000, \ldots\}$. Given two strings p, q, pq is the concatenation of p and q.
2. A subset $S \subseteq X^*$ is prefix-free iff there are no nonempty strings p, q such that $p, pq \in S$; that is, whenever the string $pq \in S$ then its prefix p is not in S.
3. An r.e. prefix code S is an r.e. prefix-free subset of X^*.
4. An r.e. prefix code S is a prefix-free r.e. subset of X^*.
5. A machine U is a partial-recursive function from X^* to X^*.
6. The description complexity $C_U(x)$ based on U is $C_U(x) = \min\{|p| : U(p)\downarrow = x\}$. U is called universal if for every further machine V there is a constant c with $\forall x\,[C_U(x) \leq C_V(x) + c]$.
7. If the domain of U is prefix-free, one also writes $H_U(x)$ for $\min\{|p| : U(p)\downarrow = x\}$ and says that U is universal iff for every further machine V with *prefix-free domain* there is a constant c with $\forall x\,[H_U(x) \leq H_V(x) + c]$.

A basic result of algorithmic information theory says that such universal machines exist [1, 11]. Here some examples for prefix-free machines: Given a sequence $V_0, V_1, V_2, \ldots$ of all prefix-free machines, one can define that $U_{ad}(1^n 0x) = V_n(x)$ for all n and $x \in \mathrm{dom}(V_n)$; then U_{ad} is a universal machine. This is the standard example and machines of this type are called "universal by adjunction". Furthermore, one can make from a given universal machine U_{gvn} a new machine U_{ev} such that the domain of U_{ev} only contains strings of even length: the idea is to define that $U_{ev}(x0) = U_{gvn}(x)$ for all x is in the domain of U_{gvn} with odd length; $U_{ev}(x) = U_{gvn}(x)$ for all x in the domain of U_{gvn} with even length; $U_{ev}(x)$ is undefined for all other x. Assuming that $\sum_n 2^{-H(n)} < 1/2$, Figueira, Stephan and Wu [8, Proposition 3] construct a universal machine U such that for each n and each $m \geq n$ there is exactly one $x \in X^n$ with $U(x)\downarrow = n$; such a machine cannot be universal by adjunction.

In general, the underlying machine is fixed to some default and the complexities C (plain) and H (prefix-free) are written without any subscript [7]. Now a prefix code is called universal iff it is the superset of the domain of a prefix-free universal domain.

For a prefix-free set V, let Ω_V be (the set representing the binary course-of-values of the real number) $\sum_{p \in V} 2^{-|p|}$. Ω-numbers turned out to be left-r.e. (as immediate by the definition). Chaitin [4] proved that if V is the domain of a universal prefix-free machine, then Ω_V is Martin-Löf random. Here Martin-Löf random sets are those which cannot be covered by Martin-Löf randomness tests; equivalently, a set A is Martin-Löf random iff $H(A(0)A(1)A(2)\ldots A(n)) \geq n$ for almost all n. Calude, Hertling, Khoussainov and Wang [2] and Kučera and Slaman [12] showed that the converse is also true and every left-r.e. Martin-Löf random set corresponds to the halting probability of some universal machine. Later, Calude and Staiger [3] extended this work by considering the relations between domains of prefix-free machines and their r.e.

prefix-free supersets. They established basic results and showed that such supersets cannot be recursive. In the present work, their results are extended as follows:

1. Let $s_W(n)$ denote the number of strings of length n in W and $s_W(n,m) = \sum_{i=n}^{n+m} s_W(i)$. A prefix-free r.e. set W is the superset of the domain of a prefix-free universal machine iff there is a constant c such that $s_W(n,c) \geq 2^{n-H(n)}$ for all n.
2. A prefix-free r.e. set W is the domain of some universal machine iff there exists a constant c such that $H(\langle n, s_W(n,c)\rangle) \geq n$ for all n.
3. There are prefix-free r.e. sets which satisfy the second but not the first condition; an example is any prefix-free r.e. set which has for almost all n that $s_W(n) = 2^{n-H(n)}$.
4. If W is an r.e. prefix-free superset of the domain of a universal machine U, then Ω_U is Solovay reducible to Ω_W, Ω_W is Martin-Löf random and W is wtt-complete.

To some extend, these results transfer to plain universal machines and their supersets as well.

1. An r.e. set W is a superset of the domain of a plain universal machine iff there is a constant c with $s_W(n,c) \geq 2^n$ for all n.
2. An r.e. set W is the domain of a plain universal machine iff there is a constant c with $C(s_W(n,c)) \geq n$ for all n.

Furthermore, the question is investigated when an r.e. but not necessarily prefix-free set is a superset of the domain of a universal prefix-free machine. In particular the following natural question remained open: Is the domain of every plain universal machine the superset of the domain of some prefix-free universal machine?

The reader should be reminded of the following additional notions used in this paper.

The ordering $\leq_{\text{qlex}}$ is called the quasi-lexicographical, length-lexicographical or military ordering of X^*: $\lambda <_{\text{qlex}} 0 <_{\text{qlex}} 1 <_{\text{qlex}} 00 <_{\text{qlex}} 01 <_{\text{qlex}} 10 <_{\text{qlex}} 11 <_{\text{qlex}} 000 <_{\text{qlex}} 001$ and so on. Furthermore, the sets of natural numbers $\mathbb{N}$ and strings X^* are identified by saying that $n \in \mathbb{N}$ represents the unique string x with $\#\{y \in X^* : y <_{qlex} x\} = n$. This is in particular useful in order to extend concepts like complexity to natural numbers without defining these concepts twice.

The function $a, b \mapsto \langle a, b\rangle$ is Cantor's pairing function of a and b: $\langle a, b\rangle = (a+b)(a+b+1)/2+b$.

A real number q is Solovay reducible to a real number r if there is an infinite approximation $a_0, a_1, a_2, \ldots$ of q from below such that there is some positive real constant $c > 0$ and some recursive approximation $b_0, b_1, b_2, \ldots$ of r from below such that $(a_{a+1} - a_s)c > b_{s+1} - b_s$ for all s. Similarly a set A is Solovay reducible to B if $\sum_{n\in A} 2^{-n}$ is Solovay reducible to $\sum_{n\in B} 2^{-n}$ as real numbers.

Further unexplained notation can be found in the books of Odifreddi [15], Calude [1] and Li and Vitányi [11].

2 Universal r.e. prefix codes

Recall that a *prefix-free universal machine* U is a prefix-free machine such that for every further machine V there is a constant c such that for every $p \in \text{dom}(V)$ there is a $q \in \text{dom}(U)$ with

$U(q) = V(p) \wedge |q| \leq |p| + c$. Following [3], a universal r.e. prefix code $A \subset X^*$ is an r.e. prefix-free set containing the domain of a prefix-free universal machine. The major goal of this section is to clarify the relation between domains of prefix-free universal machines and universal r.e. prefix codes.

For every $V \subset X^*$, let the spectrum function $s_V : X^* \to \mathbb{N}$ be defined as $s_V(n) = \#(V \cap X^n)$ and $s_V(n,m) = \sum_{i=n}^{n+m} s_V(i)$. Furthermore, for machines U, $s_U(n)$ is just $s_{\mathrm{dom}(U)}(n)$.

Theorem 1. *If U is a universal prefix-free machine then there exists a constant c such that $H(\langle n, s_U(n,c)\rangle) \geq n$ for all n.*

Proof. Assume by way of contradiction that this fails. Now choose c to be a multiple of 3 such that:

1. for every $p \in \mathrm{dom}(U)$ there is a $q \in \mathrm{dom}(U)$ with $|q| < |p| + c/3$ and $U(q) >_{qlex} U(p)$;
2. for every $p \in \mathrm{dom}(U)$, $H(U(p)) < H(p) + c/3$;
3. $H(p_n) \leq H(\langle n, s_U(n,c)\rangle) + c/3$, where p_n is the quasi-lexicographically smallest string in $\mathrm{dom}(U)$ such that $n \leq |p_n| \leq n + c$ and $U(p_n) \geq_{\mathrm{qlex}} U(q)$ for all $q \in \mathrm{dom}(U)$ with $n \leq |q| \leq n + c$.

Note that the third condition can be satisfied as there is a three-place partial-recursive function with inputs m, n and c with the following properties: this function simulates U until U has halted on a set R of m strings r with $n \leq |r| \leq n + c$ and it then outputs the length-lexicographic first $r' \in R$ for which $U(r')$ is length-lexicographically maximal: $U(r') \geq_{\mathrm{qlex}} U(r)$ for all $r \in R$. The function terminates and outputs p_n in the case that $m = s_U(n,c)$. Now the complexity of the output of this two-place function is bounded by $H(\langle n, s_U(n,c)\rangle) + 2\log(c) + c'$, for some constant c' and hence for all sufficiently large c the third condition is satisfied.

Note that by the first item it holds that $U(q) <_{\mathrm{qlex}} U(p_n)$ for all $q \in \mathrm{dom}(U)$ with $|q| \leq n + 2c/3$. Hence $|p_n| \geq n + 2c/3$. By the second item it holds that $H(p_n) \geq n + c/3$. By the third item it then follows that $H(\langle n, s_U(n,c)\rangle) \geq n$. □

Theorem 2. *There exists a prefix-free machine W and a universal prefix-free machine U such that $\mathrm{dom}(U) \subset \mathrm{dom}(W)$ and W is not universal.*

Proof. Let U be a universal prefix-free machine such that $\Omega_U < 1/2$. Now one can build by the Kraft-Chaitin Theorem a prefix-free set W such that for all n either $s_U(n) = s_W(n) = 0$ or there is a natural number m with $2^m \leq s_U(n) < s_W(n) = 2^{m+1}$. As $s_U(n) \leq s_W(n)$ for all n, one can make a partial-recursive one-one function f from $\mathrm{dom}(U)$ into W such that $|f(p)| = |p|$ for all $p \in \mathrm{dom}(U)$; this defines a further partial function from $f(U)$ to X^* by mapping $f(p) \mapsto U(p)$ for all $p \in \mathrm{dom}(U)$ which is a universal machine whose domain is a subset of W. It follows that W is a prefix-free superset of the domain of some universal function. Furthermore, for every constant c, the machine

$$n \mapsto H(\langle n, s_W(n,c)\rangle)$$

is logarithmic in n as for each value $s_W(m)$ has only $n+1$ many possible choices: either 0 or 2^m for some $m \in \{0,1,\ldots,n\}$. Hence, by Theorem 1, the set W cannot be the domain of a prefix-free universal machine. □

Although the complexity of a universal prefix code might not be large up to a given length n, the next result shows that the number

$$\Omega_W = \sum_{x \in W} 2^{-|x|}$$

is Martin-Löf random, a property shared with the domains of prefix-free universal machines. Note that there is no contradiction as for every left-r.e. real number $\rho > 0$ one can find a recursive prefix-free set W such that $\Omega_W = \rho$, see [2].

Theorem 3. *Let W be an r.e. universal prefix code. Then Ω_W is Martin-Löf random.*

Proof. Assume that U is a prefix-free universal machine whose domain is contained in the prefix-free r.e. set W. The basic idea of the proof is to show that Ω_U is Solovay reducible to Ω_W. This is done by approximating the halting probability of U such that $\Omega_{U,0} = 0$ and for every u one can compute a natural number k_u with $\Omega_{U,u+1} - \Omega_{U,u} = 2^{-k_u}$. Next one constructs a sequence $t_0, t_1, \ldots$ of integers such that there is a rational constant $\delta > 0$ with the property:

$$\forall u \left[\delta \cdot 2^{-k_u} \leq \Omega_{W,t_{u+1}} - \Omega_{W,t_u}\right].$$

This property is a reformulation of the fact that there is a Solovay-reduction from Ω_U to Ω_W. As Ω_U is Solovay-reducible to a left-r.e. set iff the latter is Martin-Löf random, the theorem follows once that δ is found [17].

The constant δ and the sequence $t_0, t_1, t_2, \ldots$ will come out of the following inductive construction: Using the Fixed-point Theorem, one can construct a r.e. prefix-free set V using a constant c such that for every $x \in V$ there is a $p \in \mathrm{dom}(U)$ with $U(p) = x \wedge |p| \leq |x| + c$. Now one defines V in stages:

1. An invariance of the construction is $\Omega_{V,u} = \Omega_{U,u}$ for all u.
2. The initialisation is $t_0 = 0$ and $V_0 = \emptyset$ which is consistent with the given invariance.
3. At stage u, assume that t_u, V_u and W_u are defined. Let k_u be the unique integer with

   $$2^{-k_u} = \Omega_{U,u+1} - \Omega_{U,u}.$$

   Find a natural number m_u which is so large that $2|W_{t_u}| < 2^{m_u}$. By the Kraft-Chaitin Theorem one can select 2^{m_u} strings of length $k_u + m_u$ which are not yet in V_u and put them as new elements into V_{u+1}. This adds 2^{-k_u} to Ω_V giving

   $$\Omega_{V,u+1} = \Omega_{V,u} + 2^{m_u} \cdot 2^{-k_u - m_u} = \Omega_{U,u} + 2^{-k_u} = \Omega_{U,u+1}.$$

   Furthermore, one can select t_{u+1} to be the first stage beyond t_u where for every string $x \in V_{u+1}$ there is an $y \in \mathrm{dom}(U_{t_{u+1}}) \cap W_{t_{u+1}}$ such that $|y| \leq |x| + c$ and $U(y) = x$; as at least half of these strings y had not been in W_{t_u} it follows that

   $$\Omega_{W,t_{u+1}} - \Omega_{W,t_u} \geq 2^{-k_u - c - 1}.$$

4. The last equation of the activity at stage u permits to choose $\delta = 2^{-c-1}$.

Hence Ω_U is Solovay reducible to Ω_W and Ω_W is Martin-Löf random [17]. □

Theorem 4. *If W is an r.e. universal prefix code then there exist two constants c, d such that*

$$\forall n \left[2^{n-H(n)-d} \leq s_W(n,c) \leq 2^{n-H(n)+d}\right]. \tag{1}$$

Proof. It is well-known that for each r.e. prefix-free set there is a constant d' such that

$$\forall n \left[s_W(n) \leq 2^{n-H(n)+d'}\right].$$

Therefore given c one can select d such that $d \geq d' + c + 2$ in order to get the inequality of the right hand side in (1). For the left hand side, let U be a universal machine with $\mathrm{dom}(U) \subseteq W$ and take c so large that $\forall n\,[H(\langle n, s_U(n,c)\rangle) \geq n]$. The prefix-free machine V codes pairs $\langle n, m\rangle$ of natural numbers in a prefix-free way: $V(p0^e1q) = \langle n, m\rangle$ if $U(p) = n$, m is the binary value of q and $|q| = n - |p| - 2e$. Thus there is a constant c_V depending on the machine V such that $H(\langle n, m\rangle) \leq n - e + 1 + c_V$ for all $m < 2^{n-|p|-2e}$.

Since $\forall n\,[H(\langle n, s_U(n,c)\rangle) \geq n]$, it follows that $s_U(n,c) \leq 2^{n-H(n)-2e}$ can hold only for $e < c_V + 1$, that is, there is a maximal value e for which there are values of n with

$$s_U(n,c) \leq 2^{n-H(n)-2e}.$$

Taking now d to be the maximum of $c + d' + 2$ from above and $2e + 2$ from the current choice of e establishes this theorem. □

If W is an r.e. universal prefix code, then one can use the constants c, d above to compute for every n the value $H(n)$ up to a constant error. It follows that one can find for every number n a number m with $H(m) > n$: one just takes that m below 4^n for which $m - \log(s_W(m,c))$ is maximal and the choice is right in all but finitely many places. Using Merkle's result on complex sets [9] or Arslanov's completeness criterion for weak truth-table reducibility in combination with the fact that W has r.e. dnr Turing degree [15], one obtains that W is wtt-complete.

Corollary 5. *If W is an r.e. universal prefix code then W is weak truth-table complete, that is, $\mathbb{K} \leq_{wtt} W$.*

The next result is the converse of Theorem 1 and had been deferred to this place as it builds on the above results. This permits to give a characterisation of the domains of prefix-free universal machines in terms of the complexity of the function $s_V(n,m)$. The constant c comes in as there are universal machines which use only programs of even length and so on.

Theorem 6. *Assume that W is an r.e. prefix-free set such that there is a constant c with $\forall n\,[H(\langle n, s_W(n,c)\rangle) \geq n]$. Then W is the domain of a universal prefix-free machine.*

Proof. Let c as fixed above. First note that there is a constant d such that

$$\forall n \ [H(\langle n, s_W(n,c)\rangle) \leq n + d] \ .$$

The reason is that there is a constant e such that

$$\forall n \ \left[s_W(n,c) \leq 2^{n-H(n)+e}\right] \ ,$$

by Theorem 4; hence one can code n with a program p having the length of $H(n)$ bits and then $s_W(n,c)$ given n with $n+e-|p|$ bits. The constant d might be a bit larger than e as one has to translate this coding into the language of the universal machine used.

Let $p_0, p_1, p_2, \ldots$ be a recursive one-one enumeration of the domain of some prefix-free universal machine U. Now one builds, using the Recursion Theorem, a recursive sequence $t_0, t_1, t_2, \ldots$ such that for some constant b the following holds for all s:

1. $H(s_{W_{t_s}}(m,c)) < |p_s| + (m + b - |p_s|)/2$ for all s and $m \geq |p_s|$.
2. For every s there is a string $q_s \in W_{t_{s+1}} - W_{t_s}$ with $|q_s| \leq| \ p_s| + b + c$.

Note that the first condition together with Theorem 1 implies that there exists a string q_s as desired in $W - W_{t_s}$. The second condition then allows us to choose t_{s+1} so large that the string q_s is actually in $W_{t_{s+1}}$.

Finally, one defines the following machine V defined on the domain of W: For any $q \in W$ find the unique s such that $q \in W_{t_{s+1}} - W_{t_s}$ and let $V(q) = U(p_s)$.

As $|q_s| \leq| \ p_s| + b + c$ and $q_s \in W_{t_{s+1}} - W_{t_s}$, it follows that $U(p_s)$ has a program at the machine V which is at most $b + c$ bits longer than p_s, hence V is a universal prefix-free machine with domain W. □

3 Plain versus prefix-free description complexity

The main result of this section is the following theorem which parallels Theorems 1, 4 and 6 in the previous section for universal plain machines. Note that X^* would be a legitimate superset of the domain of a plain universal machine in the context of this section, as there are no such requirements like prefix-freeness.

Theorem 7. *Given an r.e. set W, the equivalences* (1) $\Leftrightarrow$ (2) *and* (3) $\Leftrightarrow$ (4) *hold for the following four conditions.*
(1) *There is a constant c such that $s_W(n,c) \geq 2^n$ for all n.*
(2) *W is the superset of a domain of a plain universal machine.*
(3) *There is a constant c with $C(s_W(n,c)) \geq n$ for all n.*
(4) *W is the domain of a plain universal machine.*

Proof. (1) $\Rightarrow$ (2): One can construct, for every n which is a multiple of $c+1$ and uniformly recursive in n, a one-one mapping from $A_n = X^n \cup X^{n+1} \cup \ldots \cup X^{n+c}$ into W such that all $p \in A_n$ is mapped into $W \cap A_{n+c+1}$; these mappings just enumerate the first 2^{n+c+1} elements of

$W \cap A_{n+c+1}$ and then map those in A_n in a one-one manner into these elements. This mapping has a partial-recursive and one-one inverse f whose domain is a subset of W and whose range is the full set X^*; note that $|f(p)| \geq |p| - 2c - 2$ for all p where $f(p)$ is defined.

If U is a plain universal machine, then the mapping $p \mapsto U(f(p))$ is also a plain universal machine with its domain being a subset of W; this completes the proof for case (1).

(2) $\Rightarrow$ (1): There is a constant c such that every string of length $n+1$ has at most plain description complexity $n+c$. At least half of these strings does not have plain description complexity below n. Thus it follows that for at least half of the 2^{n+1} strings x of length $n+1$ there is a $p \in W$ with $n \leq |p| \leq n+c$ and $U(p) = x$. Thus $s_W(n,c) \geq 2^n$.

(3) $\Rightarrow$ (4): Fix the number c and follow closely the proof of Theorem 6. First note that there is a constant d such that

$$\forall n \; [C(s_W(n,c)) \leq n + d \,].$$

Let $p_0, p_1, p_2, \ldots$ be a recursive one-one enumeration of the domain of a plain universal machine U. Now one builds, using the Recursion Theorem, some recursive sequence $t_0, t_1, t_2, \ldots$ such that for some constant b the following holds for all s:

1. $C(s_{W_{t_s}}(m,c)) < |p_s| + (m + b - |p_s|)/2$, for all s and $m \geq |p_s|$.
2. For every s there is a string $q_s \in W_{t_{s+1}} - W_{t_s}$ with $|q_s| \leq |p_s| + b + c$.

Note that the first condition together with Theorem 1 imply that there exists a string q_s as desired in $W - W_{t_s}$; by virtue of the second condition one can choose t_{s+1} so large that the string q_s is actually in $W_{t_{s+1}}$.

Now the following machine V is defined on the domain of W: For any $q \in W$ find the unique s such that $q \in W_{t_{s+1}} - W_{t_s}$ and let $W(q) = U(p_s)$.

As $|q_s| \leq |p_s| + b + c$ and $q_s \in W_{t_{s+1}} - W_{t_s}$, it follows that $U(p_s)$ has a program for the machine V which is at most $b+c$ bits longer than p_s, hence V is a plain universal machine with domain W.

(4) $\Rightarrow$ (3): Let U be the universal machine with domain W. For each n, let x_n be that string in W which is enumerated last into $W \cap (X^0 \cup X^1 \cup X^2 \cup \ldots \cup X^n)$. Note that one can compute from x_n and $(n - |x_n|)/2$ a string y_n of length n which is not in W; taking s to be the first number with $x_n \in W_s$, y_n is just the length lexicographic first string of X^n which is outside the set $\{U(p) : p \in W_s \wedge |p| < n\}$. On the one hand, one has that

$$C(y_n) \leq C(x_n) + (n - |x_n|)/2 + c' \leq |x_n| + (n - |x_n|)/2 + c'',$$

for some constants c', c'' and all n; on the other hand one has that $C(y_n) \geq n$. It follows that $|x_n| \geq n - 2c''$ and $C(x_n) \geq n - c' - c''$ for all n.

Assume now by way of contradiction that for every $c > c' + c''$ there exists an n_c with $C(s_W(n_c, c)) < n_c$. Then it follows that $C(x_{n_c+c}) \leq n_c + c/2 + c'''$ for some constant c''' and all $c > c' + c''$. To see this, note that one can code this $s_W(n_c, c)$ by a string u. Furthermore, one can code x_{n_c+c} by a string of the form $a1^b0^{b'}1u$ where $a \in \{0,1\}$, $c = 2b + a$ and $b' = |n_c| - |u| > 0$.

Now one can compute a, b, b', u from $a1^b0^{b'}1u$ and has that $n_c = |u|+b'$ and $c = 2b+a$. Afterwards one can compute $s_W(n_c, c)$ from u and has that x_{n_c+c} is the string number $s_W(n_c, c)$ among those strings enumerated into W which have at least length n_c and at most length $n_c + c$. Hence, as said above, $C(x_{n_c+c}) \leq n_c + c/2 + c'''$ for some constant c''', the value of c''' depends then on the translation of the description $a1^b0^{b'}1u$ into the universal machine on which C is based. Hence $n_c + c - c' - c'' \leq n_c + c/2 + c'''$ and $c/2 \leq c' + c'' + c'''$, a contradiction to the assumption that c could take any value greater than $c' + c''$. Thus there is a $c > c' + c''$ for which n_c does not exist and it follows for this c that $\forall n\,[C(s_W(n, c)) \geq n]$. This completes the proof. □

A consequence of Theorem 7 is that the compressible strings (for the plain description complexity) form a domain of a universal machine.

Corollary 8. *Let $W = \{p \in X^* : C(p) < |p|\}$. Then there is a universal plain machine with domain W.*

Proof. Let C_s be an approximation of the complexity C from above and let U be the underlying plain universal machine. Now define a machine V on input of the form 0^i1^j0p as follows:

1. Let $n = |p| + i + 1$.
2. Determine $m = U(p)$.
3. If m is found, search for the first stage s such that there are at least m strings in the set $\{q : n \leq |q| \leq n + 2j \wedge C_s(q) < |q|\}$.
4. If m, s are found, let $V(0^i1^j0p) = r$ be the lexicographic first string of length $n + 2j$ with $C_s(r) \geq |r|$.

Note that $V(0^i1^j0p)$ is defined iff the second and third step of this algorithm terminate. There is a constant d such that

$$\forall i, j > 0\ \left[C(V(0^i1^j0p)) < i + j + |p| + d\right].$$

Let $c = 2d$ and assume by way of contradiction that there is a number n with $C(s_W(n, c)) < n$. Then there would be a p with $|p| < n$ and $U(p) = s_W(n, c)$. Let $i = n - |p| - 1$ and let $j = d$. By construction, $V(0^i1^j0p)$ is a string of length $n + c$ not in W and

$$C(V(0^i1^j0p)) \leq i + j + |p| + d = n + c - 1 < n + c\,.$$

These two facts contradict together the definitions of c, d and W. Hence W is the domain of a universal machine by Theorem 7. □

It is easy to see that the domain of a plain universal machine cannot be the subset of any prefix-free set. But the converse question is more interesting. The first theorem gives some minimum requirement on the function s_V.

Theorem 9. *Assume that V is the superset of the domain of a prefix-free universal machine. Then either there is a constant c such that $s_V(n, c) \geq 2^n$ for all n or the Turing degree of s_V is that of the halting problem.*

Proof. Let V be an r.e. superset of the domain of the universal machine U and assume that for every constant c there is a natural number n with $s_V(n,c) < 2^n$.

Now one defines a further prefix-free machine W as follows: for every $p \in dom(U)$, let t be the time the computation of $U(p)$ needs to converge and let n be the first number such that $s_{V,t}(n, 4|p|) < 2^n$. Now let $W(q) = q$ for all $q \in \{p\} \cdot X^{n+|p|}$
By definition, there is a constant c such that for every q in the domain of W there is an r in the domain of U with $U(r) = q$ and $|r| \leq |\, q| + c$. It follows that $s_U(n, 4|p|) \geq 2^{|p|+n} - 2^n \geq 2^n$ for all $p \in \mathrm{dom}(U)$ with $|p| > c$. Hence there is a string of length up to $4|p| + n$ in $V - V_s$.

Now $\mathrm{dom}(U) \leq_T V$ by the following algorithm: on input p, search the first n such that $s_V(n, 4|p|) < 2^n$. This number exists by assumption on V. Then determine the time t such that $V_t(q) = V(q)$ for all q with $|q| \leq n + 4|p|$ — this can be done easily relative to the oracle V. If $U(p)$ is defined within t steps then output "$p \in \mathrm{dom}(U)$" else output "$p \notin \mathrm{dom}(U)$". It can easily be verified that the whole knowledge needed about V is only the values of s_V and $s_{V,t}$, hence one has even that $\mathrm{dom}(U) \leq_T s_V$. □

Note that for each constant c the set $\{0^c p : |p| \text{ is a multiple of } c\}$ is a superset of the domain of some universal prefix-free machine; hence the "either-condition" Theorem 9 cannot be dropped. The next result shows that the "or-condition" is not sufficient to guarantee that some subset is the domain of a prefix-free universal machine.

Theorem 10. *Let V be an r.e. set such that for every c there is an n with $s_V(n,c) < 2^n$. Then there is an r.e. set V' with $s_V = s_{V'}$ such that V' does not contain the domain of any prefix-free universal machine.*

Proof. The central idea is to construct by induction relative to the halting problem a sequence $p_0, p_1, p_2, \ldots$ of strings such that each p_{e+1} extends p_e and $p_e \in W_e$ whenever this can be satisfied without violating the extension-condition. Furthermore, the set V' is constructed such that for each length n one enumerates $s_V(n)$ many strings of length n into V' and chooses each string $w \in X^n$ such that w is different from the strings previously enumerated into V' and one satisfies that w extends the approximations $p_{0,n}, p_{1,n}, \ldots, p_{e,n}$ of $p_0, p_1, \ldots, p_e$ for the largest possible e which can be selected.

For any fixed e it holds for almost all n that $p_{e,n} = p_e$ and that $s_V(n) \leq 2^{n-|p_e|}$ implies that all members of $V' \cap X^n$ extend p_e. By assumption there is for each constant $c > |p_e|$ a sufficiently large n such that $s_{V,4c} < 2^n$ and all members of V' of length $n + c, n + c + 1, \ldots, n + 4c$ extend p_e. Assume now that W_e is the domain of a universal machine. Then, for one of these constants c the corresponding n has in addition the property that there is a member of W_e of between length $n + c$ and $n + 2c$. If this member of W_e is not in V' then W_e is not a subset of V'. If this member of W_e is in V' then it is an extension of p_e and by the way p_e is chosen it follows that also $p_e \in W_e$, a contradiction to the assumption that W_e is prefix-free. Hence none of the W_e is a subset of V' and the domain of a prefix-free universal machine. □

The previous result is contrasted by the following example.

Example 11. *Assume that V is an r.e. set (not prefix-free) such that there is a real constant $c > 0$ with $s_V(n) \cdot 2^{-n} > c$ for all n and assume that f is a recursive function with $\sum_n 2^{-n} f(n) < c$. Then there is a prefix-free recursive subset $W \subseteq V$ with $s_W(n) = f(n)$ for all n.*

The set W can be constructed by simply picking, for $n = 0, 1, 2, 3, \ldots$, exactly $f(n)$ strings of length n out of V which do not extend previously picked shorter strings.

The main question remains which conditions on s_V guarantee that V has a subset which is the domain of a prefix-free universal machine. In the light of Theorem 10 a necessary condition is that $\exists c \forall n \, [s_V(n, c) \geq 2^n]$. One might ask whether this condition is also sufficient. By Theorem 7 this condition characterises the supersets of plain universal machines; hence one can restate the question as follows.

Open Problem 12. *Is the domain of every plain universal machine a superset of the domain of a prefix-free universal machine?*

4 Discussion

The major goal was to investigate, which prefix-free r.e. sets of strings is a universal prefix code [3], that is, a superset of the domain of a universal machine. The result is that these sets V can be characterised using the function of finite sum of the spectrum function s_V: roughly speaking, $s_V(n, c)$ has to be near to $2^{n-H(n)}$. The reason is that there are universal machines having only strings of even length and so forth. Furthermore, universal prefix codes and domains of universal machines share the property that their halting probability is Martin-Löf random. But it could also be shown that not all universal prefix codes are the domain of a universal machine: while there is a universal prefix code for which $s_V(n) = 2^{n-H(n)}$ for all n, no domain of a universal machine has this property. The reason is that for such a domain there is a constant c such that $H(s_V(n, c))$ is near to n.

Instead of using $s_V(n, c)$, one can also use $s_V(0, n)$; then the characterizations are similar. Let V be a prefix-free r.e. set and W be an r.e. set:

1. V is the domain of a plain universal machine iff there exists a natural number c such that $\forall n \, [H(\langle s_V(0, n), n) \geq n - c]$;
2. V is the superset of the domain of a prefix-free universal machine iff there exists a natural number c such that $\forall n \, [s_V(0, n) \geq 2^{n-H(n)-c}]$;
3. W is the domain of a plain universal machine iff there exists a natural number c such that $\forall n \, [C(s_W(0, n)) \geq n - c]$;
4. W is the superset of the domain of a plain universal machine iff there exists a natural number c such that $\forall n \, [s_W(0, n)) \geq 2^{n-c}]$.

A major reason that these characterizations work is that there is a constant k so that $H(\langle x, y \rangle) \leq H(x) + k$ and $C(\langle x, y \rangle) \leq C(x) + k$ for all $x \in X^*$ and $y \in \{00, 01, 10, 11\}$. This fact can be used to show that, for any universal machine U, on one hand, $s_U(n, k) \geq s_U(0, n)$. On the other hand, one can also show that $s_U(0, n) \geq s_U(n, k)/d$ for some constant $d \in \{1, 2, 3, \ldots\}$. These

ideas would then permit to prove number 1 and 3 of these equivalences. As the proofs would be essentially the same as the corresponding ones in this paper, no proofs are given here besides this short sketch of ideas.

In the case of plain description complexity, it is difficult to find the perfect analogue of those results which consider only prefix-free supersets of domains of prefix-free universal machines and therefore do not consider trivial supersets as X^*. So, in the search for a perfect analogue, one might look at the property that every r.e. prefix-free superset of the domain of a prefix-free universal machine is also the subset of such a domain. Therefore one might ask which r.e. sets are the subset of the domain of a first universal machine and the superset of the domain of a second universal machine. The answer is that these are all r.e. sets V where there is a constant c such that

$$\forall n \left[2^n \leq s_V(n,c) \leq 2^{n+c} - 2^n\right]$$

and therefore this class is not really interesting. One might consider the question whether the set is isomorphic to the prefix-free r.e. superset of the domain of a prefix-free universal machine; somehow, this question suffers already from the fact that one cannot easily find the right notion of isomorphism for this definition. Hence, although a good characterisation for the domains of universal machines had been found, the adequate question for the supersets was not found. Finding an adequate question for the case of plain description complexity may lead to further meaningful research in this direction.

A further interesting question is to characterise those r.e. sets in general which are a superset of the domain of a prefix-free universal machine. Combining of Theorem 10 with the fact that $s_U(n) \cdot 2^{-n}$ goes to 0 for n to ∞ for any prefix-free machine U, one can deduce that this characterisation cannot depend on s_V alone, but also on the way the strings are placed. It remains an interesting open problem whether every r.e. set V satisfying $\exists c \forall n \left[s_V(n,c) \geq 2^n\right]$ contains the domain of a universal prefix-free machine. Note that this question is equivalent to asking whether the domain of every plain universal machine is a superset of the domain of some prefix-free universal machine.

Furthermore, there are various definitions of universality and this paper is based on that definition where one says that U is universal if the description complexity based on U cannot be improved by more than a constant. The most prominent alternative notion says that U *is universal by adjunction* or *prefix-universal* if for every further machine V there is a finite string q such that $U(qp) = V(p)$ for all $p \in \mathrm{dom}(V)$. Universality by adjunction is quite restrictive and one cannot characterise in terms of the spectrum function s_W when a prefix-free set W is the domain of a machine which is universal by adjunction; however, this is done for normal universal machines in Theorems 1 and 6. Nevertheless, due to the more restrictive nature, prefix-free machines which are universal by adjunction have the property

$$\exists c \forall n \left[H(s_U(n)) \geq n - H(n) - c\right].$$

This property is more natural as the one in Theorem 1. Hence, it is easy to obtain machines which are universal but not universal by adjunction. An example would be a machine U obtained from V such that for all $p \in \mathrm{dom}(V)$, $U(p0) = U(p1) = V(p)$ if $|p|$ is odd and $U(p) = V(p)$ if

$|p|$ is even; it is easy to see that U inherits prefix-freeness and universality from V. Calude and Staiger [3, Fact 5] provide more information about this topic.

As the topic of the paper are mostly supersets of domains of universal machines, one could ask what can be said about the r.e. subsets of such domains. Indeed, these subsets are easy to characterise: A prefix-free r.e. set $V \subseteq X^*$ is the subset of the domain of a prefix-free universal machine iff there is a string p such that no q comparable to p is in V; an r.e. set $V \subseteq X^*$ is the subset of the domain of a plain universal machine iff there is a constant c such that $s_{X^*-V}(n, c) \geq 2^n$ for all n. Note that a subset of the domain of a prefix-free machine is also the subset of the domain of a plain universal machine, but not vice versa. Indeed, every prefix-free subset of X^* is the subset of the domain of a plain universal machine.

Acknowledgment. The authors would like to thank Wang Wei for discussions on the topic of this paper.

References

1. Cristian S. Calude. *Information and Randomness: An Algorithmic Perspective*, Second Edition, Revised and Extended, Springer-Verlag, Berlin, 2002.
2. Cristian S. Calude, Peter H. Hertling, Bakhadyr Khoussainov and Yongge Wang. Recursively enumerable reals and Chaitin Ω numbers, *Theoretical Computer Science* 255:125–149, 2001.
3. Cristian S. Calude and Ludwig Staiger. On universal computably enumerable prefix codes. *Mathematical Structures in Computer Science*, accepted.
4. Gregory J. Chaitin. *A theory of program size formally identical to information theory. Journal of the Association for Computing Machinery* 22:329–340, 1975.
5. Gregory J. Chaitin. Information-theoretic characterizations of recursive infinite strings. *Theoretical Computer Science*, 2:45–48, 1976.
6. Gregory J. Chaitin. Algorithmic information theory, *IBM Journal of Research and Development*, 21:350–359+496, 1977.
7. Rod Downey, Denis Hirschfeldt and Geoff LaForte. Randomness and reducibility. *Journal of Computer and System Sciences*, 68:96–114, 2004.
8. Santiago Figueira, Frank Stephan and Guohua Wu. Randomness and universal machines. *Journal of Complexity*, 22:738–751, 2006.
9. Bjørn Kjos-Hanssen, Wolfgang Merkle and Frank Stephan. Kolmogorov Complexity and the Recursion Theorem. STACS 2006: *Twenty-Third Annual Symposium on Theoretical Aspects of Computer Science*, Marseille, France, February 23-25, 2006. Proceedings. Springer LNCS 3884:149–161, 2006.
10. Andrei N. Kolmogorov. Three approaches to the definition of the concept "quantity of information". Problemy Peredachi Informacii 1:3-11, 1965.
11. Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Second Edition, Springer, 1997.
12. Antonín Kučera and Theodore Slaman. Randomness and recursive enumerability. *SIAM Journal on Computing* 31:199–211, 2001.

13. Per Martin-Löf. The definition of random sequences. *Information and Control* 9:602-619, 1966.
14. André Nies. *Computability and Randomness*. Oxford University Press, to appear.
15. Piergiorgio Odifreddi. *Classical Recursion Theory*. North-Holland, Amsterdam, 1989.
16. Claus Peter Schnorr. Process complexity and effective random tests. *Journal of Computer and System Sciences* 7:376–388, 1973
17. Robert Solovay. *Draft of paper on Chaitin's work.* Unpublished notes, 215 pages, 1975.