PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 00, Number 0, Pages 000–000
S 0002-9939(XX)0000-0

TWO NOTES ON SUBSHIFTS

JOSEPH S. MILLER

(Communicated by Julia Knight)

Abstract. We prove two unrelated results about subshifts. First, we give a condition on the lengths of forbidden words that is sufficient to guarantee that the corresponding subshift is nonempty. The condition implies that, for example, any sequence of binary words of lengths $5, 6, 7, \dots$ is avoidable. As another application, we derive a result of Durand, Levin and Shen [2, 3] that there are infinite sequences such that every substring has high Kolmogorov complexity. In particular, for any $d < 1$, there is a $b \in \mathbb{N}$ and an infinite binary sequence X such that if τ is a substring of X, then τ has Kolmogorov complexity greater than $d\,|\tau| - b$.

The second result says that from the standpoint of computability theory, any behavior possible from an arbitrary effectively closed subset of $n^{\mathbb{N}}$ (i.e., a Π^0_1 class) is exhibited by an effectively closed subshift. In technical terms, every Π^0_1 Medvedev degree contains a Π^0_1 subshift. This answers a question of Simpson [10].

1. Preliminaries

We use $\mathbb{N} = \{0, 1, 2, \dots\}$ for the natural numbers. For $n > 0$, let $n^{\mathbb{N}}$ denote the set of infinite sequences over the alphabet $n = \{0, \dots, n-1\}$. We write $n^{<\mathbb{N}}$ for the set of finite strings (or *words*) over the same alphabet and use λ for the empty string. The length of $\sigma \in n^{<\mathbb{N}}$ is $|\sigma|$. If $X \in n^{\mathbb{N}}$, we write $X \upharpoonright m$ for the initial segment of X of length $m \in \mathbb{N}$.

Fix $n > 1$. For $S \subseteq n^{<\mathbb{N}}$, we say that $X \in n^{\mathbb{N}}$ *avoids* S if no $\sigma \in S$ is a substring of X. The class $\mathcal{Q}_S \subseteq n^{\mathbb{N}}$ of all sequences that avoid S is called a *subshift* (or a *shift space*) and the elements of S are called *forbidden words*. See Lind and Marcus [5] for an introduction to subshifts.

Section 3 uses notions from *computability theory* (also called *recursion theory*). An introduction to computability theory can be found in the first part of Soare [11] or in the first chapter of Nies [7]. We quickly review the definitions that we need. A sequence $X \in n^{\mathbb{N}}$ is *computable* if, viewing X as a function $\mathbb{N} \to n$, there is an algorithm (or computer program) that implements X. A subset of $n^{<\mathbb{N}}$ is *computably enumerable* (*c.e.*) if there is an algorithm that lists its elements (in no particular order). If $W \subseteq n^{<\mathbb{N}}$ is c.e., then the set $\mathcal{P} \subseteq n^{\mathbb{N}}$ of sequences with no initial segment in W is called a Π^0_1 *class*. Note that every Π^0_1 class is closed with

Received by the editors August 15, 2011.
2010 *Mathematics Subject Classification.* Primary 37B10, 03D30; Secondary 03D32, 68Q30.
The author was supported by the National Science Foundation under grants DMS-0945187 and DMS-0946325, the latter being part of a Focused Research Group in Algorithmic Randomness.

respect to the product topology on $n^{\mathbb{N}}$; they should be viewed as the "effective closed" subsets.

Given $X, Y \in n^{\mathbb{N}}$, we want to define what it means for Y to be *computable from* X. Assume that $\Psi \subseteq n^{<\mathbb{N}} \times n^{<\mathbb{N}}$ is c.e. We call Ψ a *Turing functional* if whenever $\langle \sigma_0, \tau_0 \rangle, \langle \sigma_1, \tau_1 \rangle \in \Psi$ and σ_0 and σ_1 are compatible (i.e., one is an initial segment of the other), τ_0 and τ_1 are compatible. For any $X \in n^{\mathbb{N}}$, we define $\Psi(X) = \bigcup_{n \in \mathbb{N}} \bigcup_{\langle X \upharpoonright n, \tau \rangle \in \Psi} \tau$ (where a union of compatible strings is the shortest string or sequence having all of them as initial segments). If $\Psi(X)$ is an infinite sequence, we say that $\Psi(X)$ *converges* and that $Y = \Psi(X)$ is (*Turing*) *computable* from X. Note that Y is computable from X exactly if there is an algorithm that implements Y given a "black box" implementation of X.

Let $\mathcal{P}, \mathcal{Q} \subseteq n^{\mathbb{N}}$. We say that $\mathcal{P}$ is *Medvedev* (or *strongly*) reducible to $\mathcal{Q}$ and write $\mathcal{P} \leq_s \mathcal{Q}$ if there is a Turing functional Ψ such that if $A \in \mathcal{Q}$, then $\Psi(A) \in \mathcal{P}$ [6]. In other words, $\mathcal{P} \leq_s \mathcal{Q}$ if there is a uniform algorithm to transform any element of $\mathcal{Q}$ to an element of $\mathcal{P}$. Medvedev reducibility gives us a way to divorce the combinatorial structure of a class from the effective content of its members. As expected, we write $\mathcal{P} \equiv_s \mathcal{Q}$ to mean that $\mathcal{P} \leq_s \mathcal{Q}$ and $\mathcal{Q} \leq_s \mathcal{P}$ and call the equivalence classes *Medvedev degrees*.

Note that the least Medvedev degree, referred to as *zero*, consists exactly of the classes that contain computable sequences. To see this, assume that $X \in \mathcal{P}$ is computable. Consider the Turing functional Ψ containing $\langle \sigma, \tau \rangle$ exactly when τ is a prefix of X. Since $\Psi(A) = X \in \mathcal{P}$, for all A, this Ψ witnesses the fact that $\mathcal{P}$ is Medvedev reducible to every other class. On the one hand, everything below a class with a computable member must also have a computable member. So anything in the least degree must have this property.

See Rogers [8] for other basic facts about the Medvedev degrees.

2. A sufficient condition for a subshift to be nonempty

The fact that there is an explicit sequence of lengths such that any sequence of binary words with those lengths gives a nonempty subshift was proved by Cenzer, Dashti and King [1, Theorem 3.2]. The sequence of lengths they give is $6 \cdot 2^{2i(i+3)}$, for $i \in \mathbb{N}$. While they had no need to be efficient, it is interesting to note that, in fact, any sequence of binary words of lengths $5, 6, 7, \ldots$ is avoidable. This follows from a simple condition on the lengths of strings in $S \subseteq n^{<\mathbb{N}}$ that guarantees that $\mathcal{Q}_S$ is nonempty.

Proposition 2.1. *Let $S \subseteq n^{<\mathbb{N}}$. If $\lambda \notin S$ and there is a $c \in (1/n, 1)$ such that*

$$\sum_{\tau \in S} c^{|\tau|} \leq nc - 1,$$

then there is an $X \in n^{\mathbb{N}}$ that avoids S.

Proof. Fix $c \in (1/n, 1)$. Let $p = \sum_{\tau \in S} c^{|\tau|}$ and assume that $p \leq nc - 1$. For each $\sigma \in n^{<\mathbb{N}}$, let $w(\sigma) = \sum_{\tau \in S} \sum \{c^{|\rho|} \colon |\rho| < |\tau| \text{ and } \sigma\rho \text{ ends in } \tau\}$. Think of $w(\sigma)$ as a measure of the pending threats to an extension of σ ending in an element of S. Note that $w(\lambda) = 0$. Our goal is to build a sequence $X \in n^{\mathbb{N}}$, one digit at a time, so that the weight stays below 1. Such an X avoids S because as long as $w(\sigma) < 1$, we know that σ itself does not end in an element of S. Say that we currently have σ such that $w(\sigma) < 1$. It is not hard to see that $\sum_{0 \leq i < n} w(\sigma i) = w(\sigma)/c + p/c$. This

is because every threat counted on the left must either have been a pending threat to σ (in which case its weight has gone up by a factor of $1/c$) or it is a new threat (and the weight of all possible new threats is p/c). Since $w(\sigma) + p < 1 + p \leq nc$, we have $\sum_{0 \leq i < n} w(\sigma i) < n$. Therefore, $w(\sigma i) < 1$ for some $0 \leq i < n$, allowing us to continue the construction. □

Note that $c > 1/n$ is not used in the proof, only that $c > 0$. But if S is nonempty and $c > 0$, then $nc - 1$ must be positive for the hypothesis to hold. Therefore, $c > 1/n$ is not an additional restriction. Also note that if $n = 2$, then we do not have to explicitly assume that $\lambda \notin S$ in the result above. This is because $2c - 1 < 1$, so the bound on the sum already implies that S does not contain the empty string.

The condition given in Proposition 2.1 does not characterize avoidability. This is hardly surprising because it is only a condition on the lengths (with multiplicity) of the members of S, and whether S is avoidable depends on more than just the lengths of its elements. For example, $S_0 = \{00, 11, 10\}$ is not avoidable, but $S_1 = \{01, 11, 10\}$ is avoided by $0^{\mathbb{N}}$.

The next natural question is whether the condition characterizes the multisets of lengths that *guarantee* avoidability. Unfortunately, this is not the case; there is a multiset of lengths that *does not* satisfy the condition but such that every realization is avoidable. Specifically, any set S consisting of two binary strings of length 2 is avoidable, but there is no $c \in (1/2, 1)$ such that $2c^2 \leq 2c - 1$. Both claims are easily verified.

Corollary 2.2. *Assume that $S \subseteq n^{<\mathbb{N}}$ contains at most one string of each length and let $L = \{|\sigma| : \sigma \in S\}$. If*

(a) *$n = 2$ and $L \subseteq \{5, 6, 7, \dots\}$, or*
(b) *$n = 2$ and $L \subseteq \{4, 6, 8, \dots\}$, or*
(c) *$n = 3$ and $L \subseteq \{2, 3, 4, \dots\}$, or*
(d) *$n = 4$ and $L \subseteq \{1, 2, 3, \dots\}$,*

then there is an $X \in n^{\mathbb{N}}$ that avoids S.

Proof. In (a) and (b) we can apply the proposition with $c = \frac{\sqrt{5}-1}{2}$, the inverse of the golden ratio. For (c) and (d) we can use $c = 1/2$. Also note that (d) follows easily from (c) because a string of length 1 simply eliminates a digit. □

Remark 2.3. One might wonder if the corollary could be improved to allow lengths $L \subseteq \{4, 5, 6, \dots\}$ when $n = 2$. We can refute this with an example. Let

$$S = \{1000, 10011, 100101, 1011111, 10111101, 101110101,$$
$$1011101101, 10110101101, 101101010101, 1011010101101,$$
$$0^m, 1^m, (100)^m, (1110)^m, (110)^m, (10)^m\},$$

where $m \in \mathbb{N}$ is large. Assume that $X \in 2^{\mathbb{N}}$ avoids S. Because $0^m, 1^m \in S$, we know that X contains a substring of the form 10. By taking a tail of X, we can assume without loss of generality that it begins with 10. Since X avoids 1000, no more than two zeros occur together. If 1001 is a substring of X, then because $10011, 100101 \in S$, we know that X must end in $(100)^{\mathbb{N}}$. But X avoids $(100)^m$, so this is impossible. Therefore, X avoids 1001. As a consequence, it avoids 100; all zeros in X occur in isolation.

Since $1011111, 10111101 \in S$, no more than three ones occur in succession in X. But X avoids 101110101 and 1011101101, so if 111 ever occurs, then X ends in $(1110)^{\mathbb{N}}$. This is impossible because $(1110)^m \in S$. Therefore, no more than two ones occur together in X. Now since S also contains 10110101101, 101101010101 and 1011010101101, if 11 is ever a substring of X, then X must end in $(110)^{\mathbb{N}}$. But $(110)^m \in S$ eliminates this possibility, so all ones in X occur in isolation too. Therefore $X = (10)^{\mathbb{N}}$, contradicting the fact that $(10)^m \in S$.

The next application of Proposition 2.1 involves *prefix-free* (*Kolmogorov*) *complexity*, a notion from effective randomness. Prefix-free complexity is a function $K\colon 2^{<\mathbb{N}} \to \mathbb{N}$ that measures, in a certain sense, the complexity of finite binary strings. The intended interpretation is that τ contains $K(\tau)$ bits of information. The only technical fact we need below is *Kraft's inequality*: $\sum_{\tau \in 2^{<\mathbb{N}}} 2^{-K(\tau)} \leq 1$. For an introduction to Kolmogorov complexity, see Li and Vitányi [4] or Nies [7].

Schnorr proved that $X \in 2^{\mathbb{N}}$ is *Martin-Löf random* if and only if $K(X \upharpoonright n) \geq n - O(1)$. Almost every infinite sequence is Martin-Löf random, so we know that almost every sequence's initial segments have high complexity. What about the complexity of substrings? Here things look different; almost every infinite sequence has arbitrarily long runs of zeros, which have low complexity. Even so, Durand, Levin and Shen [2, 3] showed that we can find sequences such that every substring has *fairly* high complexity. We give another proof of their result. (See also Rumyantsev and Ushakov [9], who give a proof using the Lovász local lemma.)

Note that we are limited by the fact that if X avoids *any* string $\tau \in 2^{<\mathbb{N}}$, then $\limsup K(X \upharpoonright n)/n < 1$. In other words, the following result would fail for $d = 1$.

Corollary 2.4 (Durand, Levin and Shen [2, 3])**.** *Let $d < 1$. There is an $X \in 2^{\mathbb{N}}$ such that if $\tau \in 2^{<\mathbb{N}}$ is a substring of X, then $K(\tau) > d\,|\tau| - O(1)$.*

Proof. Fix $d \in (0,1)$ and let $b = -\log(1-d) + 1$ (where log denotes the base 2 logarithm). Let $S = \{\tau \in 2^{<\mathbb{N}} \colon K(\tau) \leq d\,|\tau| - b\}$. To apply Proposition 2.1, we let $c = 2^{-d}$. Then

$$\sum_{\tau \in S} c^{|\tau|} = \sum_{\tau \in S} 2^{-d|\tau|} \leq \sum_{\tau \in S} 2^{-K(\tau)-b} \leq 2^{-b} \sum_{\tau \in 2^{<\mathbb{N}}} 2^{-K(\tau)} \leq 2^{-b},$$

where the last step is Kraft's inequality. It is easy to show that $2^{-b} = (1-d)/2 < 2^{1-d} - 1 = 2c - 1$, for $d \in (0,1)$. □

Prefix-free complexity is not the only form of Kolmogorov complexity, and in fact, it is not the version originally, and independently, introduced by Solomonoff and Kolmogorov. But the common variants of Kolmogorov complexity differ from $K(\sigma)$ by $O(\log |\sigma|)$, and Corollary 2.4 is clearly true for any such variant.

Finally, we note that Rumyantsev and Ushakov [9] derived Corollary 2.4 from a closely related statement that does not mention Kolmogorov complexity. Their result also follows from Proposition 2.1.

Corollary 2.5 (Rumyantsev and Ushakov [9])**.** *Fix $\alpha \in [0,1)$. There is a $d \in \mathbb{N}$ such that if $S \subseteq n^{<\mathbb{N}}$ contains at most $n^{\alpha m}$ strings of length m, for each $m \geq d$, and none of length less than d, then there is an $X \in n^{\mathbb{N}}$ that avoids S.*

Proof. Fix $\beta \in (\alpha, 1)$ and let $c = n^{-\beta}$. Then $cn - 1 > 0$ and $\sum_{m \in \mathbb{N}} n^{\alpha m} c^m = \sum_{m \in \mathbb{N}} n^{(\alpha-\beta)m}$ converges. So if $d \in \mathbb{N}$ is sufficiently large, then $\sum_{m \geq d} n^{\alpha m} c^m < cn - 1$. Therefore, we can apply Proposition 2.1. □

3. Every Π^0_1 Medvedev degree contains a Π^0_1 subshift

Simpson [10] proved that every Π^0_1 Medvedev degree contains a 2-dimensional subshift of *finite type*, i.e., one for which the set of forbidden 2-dimensional words is finite. This does not hold for 1-dimensional subshifts, the type considered in the present paper. Every nonempty (1-dimensional) subshift of finite type contains periodic sequences, so they all have Medvedev degree zero. On the other hand, Cenzer, Dashti and King [1] produced a Π^0_1 subshift that contains no computable sequences, hence has nonzero Medvedev degree. Simpson [10] asked if every Medvedev degree containing a Π^0_1 class actually contains a Π^0_1 subshift. We give a positive answer.

If S is a computably enumerable (c.e.) set, then $\mathcal{Q}_S$ is a Π^0_1 class. Conversely, it is not hard to see that if $\mathcal{Q}$ is a Π^0_1 subshift, then the set S of all strings that appear in no element of $\mathcal{Q}$ is c.e. This is because S is also the set of strings that cannot be extended to an element of $\mathcal{Q}$. Since $\mathcal{Q} = \mathcal{Q}_S$, we see that Π^0_1 subshifts are exactly the subshifts defined by c.e. sets of forbidden words. We will show that from a computability-theoretic perspective, Π^0_1 subshifts can exhibit all of the behavior possible from arbitrary Π^0_1 subclasses of $n^{\mathbb{N}}$. By coding elements of $n^{\mathbb{N}}$ in binary, it is easy to see that every Π^0_1 subclass of $n^{\mathbb{N}}$ is Medvedev equivalent to a Π^0_1 subclass of $2^{\mathbb{N}}$. So for what follows, we restrict ourselves to binary sequences.

Proposition 3.1. *If $\mathcal{P}$ is a Π^0_1 class, then there is a Π^0_1 subshift $\mathcal{Q}$ such that $\mathcal{P} \equiv_s \mathcal{Q}$.*

Proof. The key feature of the coding we use is that a sequence $Y \in \mathcal{P}$ is coded by another sequence X in such a way that every tail of X also codes Y, where the method of decoding does not depend on Y or on which tail of X we are given. To do this, we code every bit of Y throughout all of X. First, we define collections of strings $\{a_\sigma\}_{\sigma\in 2^{<\mathbb{N}}}$ and $\{b_\sigma\}_{\sigma\in 2^{<\mathbb{N}}}$. Let $a_\lambda = 0$ and $b_\lambda = 1$. For $\sigma \in 2^{<\mathbb{N}}$, let

$$a_{\sigma 0} = b_\sigma a_\sigma a_\sigma, \qquad b_{\sigma 0} = b_\sigma a_\sigma a_\sigma a_\sigma,$$
$$a_{\sigma 1} = a_\sigma b_\sigma b_\sigma, \text{ and} \qquad b_{\sigma 1} = a_\sigma b_\sigma b_\sigma b_\sigma.$$

Define $S_\sigma = \{a_\sigma a_\sigma a_\sigma a_\sigma, b_\sigma b_\sigma b_\sigma b_\sigma, a_\sigma a_\sigma b_\sigma b_\sigma, b_\sigma b_\sigma a_\sigma a_\sigma, a_\sigma b_\sigma a_\sigma b_\sigma, b_\sigma a_\sigma b_\sigma a_\sigma\}$ and let $S = \bigcup_{\sigma\in 2^{<\mathbb{N}}} S_\sigma$. It is not hard to see that if $X \in 2^{\mathbb{N}}$ avoids S_λ, then (except for at most three initial bits) X is either formed from a_0 and b_0 or from a_1 and b_1. In other words, there is a *unique* way to decompose X (ignoring at most three initial bits) into a sequence of a_0's and b_0's *or* a sequence of a_1's and b_1's, but not both. Moreover, if a tail of X is formed from a_0 and b_0, then there is no occurrence of a_1 or b_1, even using the initial bits. The same holds with 0 and 1 reversed.

Let us assume, without loss of generality, that X can be decomposed as a sequence of a_0's and b_0's. Because X avoids S_0, the same analysis shows that (except for at most three initial bits *and* at most three copies of either a_0 or b_0) X is either formed from a_{00} and b_{00} or from a_{01} and b_{01}. In the first case, there is no occurrence of a_{01} or b_{01} in X, and similarly for the second case.

In this way, we can see by induction that if X avoids S, then for each $n \in \mathbb{N}$ there is a unique $\sigma \in 2^n$ such that X is formed from a_σ and b_σ (again, possibly disregarding an initial segment) and no a_τ or b_τ can occur in X unless σ and τ are comparable. On the other hand, for any $Y \in 2^{\mathbb{N}}$, the infinite sequence $\Psi(Y) = \bigcup_{n>0} a_{Y\upharpoonright n}$ avoids S. To see that $\Psi(Y)$ is well defined we observe that a_σ is a prefix of both $a_{\sigma 0}$ and $a_{\sigma 1}$ as long as $\sigma \neq \lambda$. The latter is immediate. For the former, note that $\sigma \neq \lambda$ implies that a_σ is a prefix of b_σ and hence of $a_{\sigma 0}$.

Let $W \subseteq 2^{<\mathbb{N}}$ be a c.e. set of strings such that $\mathcal{P}$ is exactly the set of sequences with no initial segment in W. Let $T = S \cup \{a_\sigma : \sigma \in W\}$. Then T is a c.e. set of strings, so the induced subshift $\mathcal{Q}_T$ is a Π^0_1 class. Moreover, Ψ is an effective reduction of elements of $\mathcal{P}$ to elements of $\mathcal{Q}_T$; hence $\mathcal{Q}_T \leq_s \mathcal{P}$. For the other direction, let $\Phi(X) = \bigcup\{\sigma \in 2^{<\mathbb{N}} : a_\sigma \text{ is a substring of } X\}$ and assume that $\Phi(X)$ stops converging as soon as an incompatibility is found. Clearly Φ is total on any X that avoids S. If we additionally assume that X avoids T, then $\Phi(X) \in \mathcal{P}$. Thus $\mathcal{P} \leq_s \mathcal{Q}_T$. □

References

1. Douglas Cenzer, S. Ali Dashti, and Jonathan L. F. King, *Effective symbolic dynamics*, Proceedings of the Fourth International Conference on Computability and Complexity in Analysis (CCA 2007) (Ruth Dillhage, Tanja Grubba, Andrea Sorbi, Klaus Weihrauch, and Ning Zhong, eds.), Electronic Notes in Theoretical Computer Science, vol. 202, Elsevier, 2008, CCA 2007, Siena, Italy, June 16–18, 2007, pp. 89–99.
2. Bruno Durand, Leonid Levin, and Alexander Shen, *Complex tilings*, Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing (New York), ACM, 2001, pp. 732–739 (electronic). MR MR2120376
3. Bruno Durand, Leonid A. Levin, and Alexander Shen, *Complex tilings*, J. Symbolic Logic **73** (2008), no. 2, 593–613. MR MR2414467 (2009f:52046)
4. Ming Li and Paul Vitányi, *An introduction to Kolmogorov complexity and its applications*, third ed., Texts in Computer Science, Springer, New York, 2008. MR MR2494387 (2010c:68058)
5. Douglas Lind and Brian Marcus, *An introduction to symbolic dynamics and coding*, Cambridge University Press, Cambridge, 1995. MR MR1369092 (97a:58050)
6. Yu. T. Medvedev, *Degrees of difficulty of the mass problem*, Doklady Akademii Nauk SSSR **104** (1955), 501–504. MR 17,448b
7. André Nies, *Computability and randomness*, Oxford Logic Guides, vol. 51, Oxford University Press, Oxford, 2009. MR MR2548883
8. Hartley Rogers, Jr., *Theory of recursive functions and effective computability*, second ed., MIT Press, Cambridge, MA, 1987. MR 886890 (88b:03059)
9. A. Yu. Rumyantsev and M. A. Ushakov, *Forbidden substrings, Kolmogorov complexity and almost periodic sequences*, STACS 2006, Lecture Notes in Comput. Sci., vol. 3884, Springer, Berlin, 2006, pp. 396–407. MR MR2249384 (2007c:68119)
10. Stephen G. Simpson, *Medvedev degrees of 2-dimensional subshifts of finite type*, to appear.
11. R. I. Soare, *Recursively enumerable sets and degrees*, A study of computable functions and computably generated sets, Perspectives in Mathematical Logic, Springer-Verlag, Berlin, 1987. MR 88m:03003

Joseph S. Miller, Department of Mathematics, University of Wisconsin, Madison, WI 53706-1388, USA

E-mail address: jmiller@math.wisc.edu