The Frequent Paucity of Trivial Strings

Jack H. Lutz
Department of Computer Science
Iowa State University
Ames, IA 50011, USA
lutz@cs.iastate.edu

Abstract

A 1976 theorem of Chaitin can be used to show that arbitrarily dense sets of lengths n have a *paucity of trivial strings* (only a bounded number of strings of length n having trivially low plain Kolmogorov complexities). We use the probabilistic method to give a new proof of this fact. This proof is much simpler than previously published proofs, and it gives a tighter paucity bound.

1 Background

A string of binary data is *trivial* if, like a string of all zeros, it contains negligible information beyond that implicit in its length. This notion of triviality has been made precise in several different ways, and these have been useful in the foundations of Kolmogorov complexity [6], information-theoretic characterizations of decidability and polynomial-time decidability [2, 8], formal language theory [4], and the theory of K-trivial sequences [7, 3].

These applications share several common features. Each uses some version of Kolmogorov complexity to quantify the information content of a string. Each *parametrizes* its triviality notion by a nonnegative integer c, defining a string to be *c-trivial* if its information content is within c bits of a triviality criterion. Most crucially, the key to each of these applications is a *paucity theorem*, stating that there are many lengths n at which there is a paucity (at most a fixed multiple of 2^c) of c-trivial strings of length n.

The first such paucity theorem, reported in 1969, was proved by Meyer [6]. Chaitin subsequently strengthened Meyer's proof, slightly relaxing his triviality notion and obtaining the following.

Theorem 1 (Chaitin [2])**.** *There is a constant $a \in \mathbb{N}$ such that, for all $n, d \in \mathbb{N}$, at most 2^{d+a} strings $x \in \{0,1\}^n$ satisfy $C(x) \leq d + C(n)$.*

Here $C(x)$ is the *plain Kolmogorov complexity* of x, the minimum number of bits required to program a fixed universal Turing machine to print the string x, and $C(n) = C(s_n)$, where $s_0, s_1, \ldots$ is a standard enumeration of $\{0,1\}^*$. (Thorough treatments of $C(x)$ appear in [5, 7, 3].)

This note concerns paucity theorems involving $\log n$, rather than $C(n)$, as a triviality criterion. Since $C(n)$ is usually close to $\log n$, one such paucity theorem can be derived from Theorem 1, as we now show.

Logarithms here are base-2. We will use the (*Schnirelmann*) *density* of a set $L \subseteq \mathbb{N}$, which is

$$\sigma(L) = \inf \left\{ \frac{|L_{<m}|}{m} \;\middle|\; m \in \mathbb{Z}^+ \right\},$$

where we write $L_{<m} = L \cap \{0, \ldots, m-1\}$ [9]. Intuitively the condition $n \in L$ holds *frequently* if $\sigma(L) > 0$. This is clearly a stronger condition than the assertion that L is infinite. To relate the triviality criteria $\log n$ and $C(n)$, define the set

$$L(r) = \left\{ n \in \mathbb{N} \;\middle|\; C(n) + r \geq \log n \right\}$$

for each $r \in \mathbb{N}$.

Observation 2. For each $R \in \mathbb{N}$, $\sigma(L(r)) \geq 1 - 2^{1-r}$.

Proof. For each $m \in \mathbb{Z}^+$, the complement $L(r)^c$ of $L(r)$ satisfies

$$\begin{aligned} (L(r)^c)_{<m} &= \{n < m \mid C(n) < (\log n) - r\} \\ &\subseteq \{n < m \mid C(n) < (\log m) - r\} \end{aligned}$$

so

$$\begin{aligned} |(L(r)^c)_{<m}| &\leq \left|\{0,1\}^{<(\log m)-r}\right| \\ &< 2^{1-r+\log m} \\ &= 2^{1-r} m. \end{aligned}$$

It follows that

$$\begin{aligned} \sigma(L(r)) &= \inf \left\{ \frac{|L(r)_{<m}|}{m} \;\middle|\; m \in \mathbb{Z}^+ \right\} \\ &\geq \inf \left\{ \frac{m - 2^{1-r} m}{m} \;\middle|\; m \in \mathbb{Z}^+ \right\} \\ &= 1 - 2^{1-r}. \end{aligned}$$

□

We now have the following easy consequence of Theorem 1.

Theorem 3 (very frequent paucity theorem)**.** *The constant a of Theorem 1 has the property that, for all $c, r \in \mathbb{N}$, the set of nonnegative integers n for which at most 2^{c+a+r} strings $x \in \{0,1\}^n$ satisfy $C(x) \leq c + \log n$ has density at least $1 - 2^{1-r}$.*

Proof. Let $a \in \mathbb{N}$ be as in Theorem 1, and let $c, r \in \mathbb{N}$. For each $n \in \mathbb{N}$, define the sets

$$B_n = \{\, x \in \{0,1\}^n | \, C(x) \leq c + \log n\}$$

and

$$B'_n = \{\, x \in \{0,1\}^n | \, C(x) \leq c + r + C(n)\}$$

and let

$$L_c = \left\{ n \in \mathbb{N} \,\middle|\, |B_n| \leq 2^{c+a+r} \right\}.$$

It suffice to show that $\sigma(L_c) \geq 1 - 2^{1-r}$.

Let $n \in L(r)$. Then $C(n) + r \geq \log n$, so $B_n \subseteq B'_n$. Applying Theorem 1 with $d = c + r$, we have $|B'| \leq 2^{c+r+a}$, whence $|B_n| \leq 2^{c+r+a}$. Hence $n \in L_c$.

We have now shown that $L(r) \subseteq L_c$. It follows by Observation 2 that

$$\sigma(L_c) \geq \sigma(L(r)) \geq 1 - 2^{1-r}.$$

□

The proofs of Theorem 1 and Meyer's earlier paucity theorem are somewhat involved. Part of this is because these early proofs were aimed at proving more, namely that

(I) for every $c \in \mathbb{N}$ there are at most 2^{c+a} *infinite* binary sequences that are *c-trivial* in the sense that *every* nonempty prefix x of such a sequence satisfies $C(x) \leq c + log|x|$; and

(II) every such c-trivial sequence is decidable.

It is clear that (I) follows immediately from Theorem 1, and it is now well understood that (II) follows directly from (I), because every isolated infinite branch of a decidable tree is decidable [3].

In the 1990s, Li and Vitanyi proved the following paucity theorem.

Theorem 4 (Li and Vitanyi [4])**.** *There is a constant $a \in \mathbb{N}$ such that, for every $c \in \mathbb{N}$, there exist infinitely many lengths n for which at most 2^{c+a} strings $x \in \{0,1\}^n$ satisfy $C(x) \leq c + \log n$.*

Theorem 4 is weaker than Theorem 3, because it only tells us that the paucity of trivial strings occurs at infinitely many lengths. Li and Vitanyi's proof of Theorem 4 is simpler than the proof of Theorem 1 (hence simplier than the proof of Theorem 3), even when one discounts the parts of the proof of Theorem 1 devoted to (I) and (II). However, even Li and Vitanyi's simplified proof is nontrivial.

2 Result

The purpose of this note is to give a *very* simple proof of a frequent paucity theorem. Our theorem's frequency condition is as strong as that of Theorem 3. However, our theorem improves on earlier paucity theorems in a significant respect: While the proofs of Theorems 1, 3, and 4 require the constant a to be as large as the number of bits required to encode a nontrivial Turing machine, our simple proof shows that it suffices to take $a = 1$.

Our simple proof has a simple intuition: As in the proof of Theorem 3, let

$$B_n = \{\, x \in \{0,1\}^n \,|\, C(x) \leq c + \log n \,\}\,.$$

We want to show that $|B_n|$ is often small. Well, the *average* of the first m values of $|B_n|$ is

$$\begin{aligned} \frac{1}{m}\sum_{n=0}^{m-1} |B_n| &= \frac{1}{m}\left|\bigcup_{n=0}^{m-1} B_n\right| \\ &\leq \frac{1}{m}\left|\{0,1\}^{<c+\log m}\right| \\ &< \frac{1}{m} 2^{c+\log m+1} \\ &= 2^{c+1}, \end{aligned}$$

so $|B_n| \leq 2^{c+1}$ must hold frequently! The details follow.

Theorem 5. *Let $c \in \mathbb{N}$.*

1 (frequent paucity). *The set of nonnegative integers n for which at most 2^{c+1} strings $x \in \{0,1\}^n$ satisfy $C(x) \leq c + \log n$ has density at least $(2^{c+1} - 1)^{-1}$.*

2 (very frequent paucity). *For every $r \in \mathbb{N}$, the set of nonnegative integers n for which at most 2^{c+r} strings $x \in \{0,1\}^n$ satisfy $C(x) \leq c + \log n$ has density at least $1 - 2^{1-r}$.*

Proof. Let $c, r \in \mathbb{N}$, and let $d = 2^{c+r}$. For each $n \in \mathbb{N}$, let

$$B_n = \{\, x \in \{0,1\}^n \,|\, C(x) \leq c + \log n \},$$

noting that $B_0 = \emptyset$, and let

$$L = \left\{ n \in \mathbb{N} \,\middle|\, |B_n| \leq d \right\}$$

Let $m \in \mathbb{Z}^+$, and let $l = |L_{<m}|$.

Consider the average

$$\mu = \frac{1}{m} \sum_{n=0}^{m-1} |B_n|.$$

We have

$$\begin{aligned} \mu &= \frac{1}{m} \left| \bigcup_{n=0}^{m-1} B_n \right| \\ &\leq \frac{1}{m} \left| \{0,\}^{<c|logm} \right| \\ &< \frac{1}{m} 2^{c+\log m+1} \\ &= 2^{c+1} \end{aligned}$$

and

$$\mu \geq \frac{1}{m}(m-l)(d+1)$$

whence

$$m \cdot 2^{c+1} > (m-l)(d+1). \qquad (*)$$

1. If $r = 1$, then $(*)$ says that

$$md > (m-l)(d+1)$$

whence

$$l > \frac{m}{d+1}.$$

Since this holds for all $m \in \mathbb{Z}^+$, it follows that $\sigma(L) \geq \frac{1}{d+1}$

2. More generally, for $r \in \mathbb{N}$, $(*)$ implies that

$$m \cdot 2^{c+1} > (m-l)2^{c+r},$$

whence

$$d > m(1 - 2^{1-r})$$

Since this holds for all $m \in \mathbb{Z}^+$, it follows that $\sigma(L) \geq 1 - 2^{1-r}$. $\square$

3 Conclusion

The simplicity of the above proof is the main contribution of this note. Its simplicity arises from its use of the first moment probabilistic method [1, 9]: Rather than deal with the cardinalities $|B_n|$ individually, it examines their average. It is an open question whether the probabilistic method can similarly simplify the proof of Theorem 1.

A brief remark on pedagogy: Li and Vitanyi's Kolmogorov complexity characterization of regular languages [4, 5] yields a simple and intuitive method for proving that languages are not regular. A possible obstacle to teaching this method in undergraduate theory courses has been that the characterization theorem relies on the (seemingly) difficult Theorem 4. The simple proof here removes that obstacle.

Acknowledgments

I thank the referees for extremely useful observations. This research was supported in part by National Science Foundation Grant 0652569. Part of this work was done during a sabbatical at Caltech and the Isaac Newton Institute for Mathematical Sciences at the University of Cambridge.

References

[1] N. Alon and J.H. Spencer. *The Probabilistic Method* (third edition). John Wiley & Sons, 2008.

[2] Gregory J. Chaitin. Information-theoretic characterizations of recursive infinite strings. *Theoretical Computer Science*, 2:45–48, 1976.

[3] Rodney G. Downey and Denis R. Hirschfeldt. *Algorithmic Randomness and Complexity*. Springer, 2010.

[4] Ming Li and Paul M.B. Vitányi. A new approach to formal language theory by Kolmogorov complexity. *SIAM Journal on Computing*, 24:398–410, 1995.

[5] Ming Li and Paul M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications* (third edition). Springer, 2008.

[6] Donald W. Loveland. A variant of the Kolmogorov concept of complexity. *Information and Control*, 15:510–526, 1969.

[7] André Nies. *Computability and Randomness*. Oxford University Press, New York, NY, USA, 2009.

[8] Pekka Orponen, Ker-I Ko, Uwe Schöning, and Osamu Watanabe. Instance complexity. *Journal of the ACM*, 41:96–121, 1994.

[9] Terence Tao and Van Vu. *Additive Combinatorics*. Cambridge University Press, 2006.