Not All Universal Source Codes Are Pointwise Universal

Tsachy Weissman*

July 28, 2004

Abstract

Let $C_n$ be a lossless code for $n$-tuples over a finite alphabet and $l_n$ denote its length function. The code sequence $\{C_n\}$ is said to be universal if for every stationary source $X = (X_1, X_2, \ldots)$

$$\frac{1}{n} EL_n(X^n) \to \overline{H}(X),$$

where $X^n = (X_1, \ldots, X_n)$ and $\overline{H}(X)$ denotes the entropy rate of $X$. It is said to be pointwise universal if for every stationary and ergodic source, with probability one,

$$\frac{1}{n} l_n(X^n) \to \overline{H}(X).$$

Pointwise universality implies universality. We show that the converse is not true by constructing a universal sequence of schemes that is not pointwise universal, hence establishing the latter as a stronger notion of universality.

Key words and phrases: Entropy rate, Lossless compression, Pointwise universality, Universal compression.

1 Introduction

A lossless source code for binary $n$-tuples is an injection from $\{0, 1\}^n$ into $\{0, 1\}^*$. Let $C_n$ be a lossless source code for binary $n$-tuples and $l_n$ denote its length function, i.e., $l_n(x^n) = |C_n(x^n)| \forall x^n \in \{0, 1\}^n$. We assume throughout $\max_{x^n} l_n(x^n) \leq cn$ for some constant $c$.

Let $\overline{H}(X)$ denote the entropy rate of a stationary source $X = (X_1, X_2, \ldots)$, i.e., $\overline{H}(X) = \lim_{n \to \infty} \frac{1}{n} H(X^n)$, where $X^n = (X_1, \ldots, X_n)$. As ascertained in [8], the entropy rate is the fundamental lower bound on attainable compression in that for any lossless code sequence and any stationary source $\liminf_{n \to \infty} \frac{1}{n} EL_n(X^n) \geq \overline{H}(X)$. It is also a lower bound in the pointwise sense that for any lossless code sequence and any stationary and ergodic source, with probability one, $\liminf_{n \to \infty} \frac{1}{n} l_n(X^n) \geq \overline{H}(X)$ (cf., e.g., [5] and references therein). The entropy rate is attainable since for any stationary process, e.g. the Shannon code satisfies $\lim_{n \to \infty} \frac{1}{n} EL_n(X^n) = \overline{H}(X)$ and, by the Shannon-McMillan-Breiman Theorem [1], satisfies also $\lim_{n \to \infty} \frac{1}{n} l_n(X^n) = \overline{H}(X)$ with probability one provided the source is also ergodic. While the Shannon code is source-dependent, there exist source codes which are universal in the following senses.

Definition 1 The sequence $\{C_n\}_{n \geq 1}$ is said to be universal if for every stationary source $X$

$$\frac{1}{n} EL_n(X^n) \to \overline{H}(X).$$

---

*Electrical Engineering Department, Stanford University, CA 94305 USA (e-mail: tsachy@stanford.edu).

1A Binary source alphabet as well as binary encoding is assumed throughout for concreteness. The results carry over trivially to the case of arbitrary finite alphabets.

2This is a benign assumption as there is no loss in even restricting attention to codes satisfying $\max_{x^n} l_n(x^n) \leq n + 1$. If a given code does not satisfy this, it can be modified, adding one bit to indicate whether we compress or not, and sending the source uncompressed whenever its associated length function is more than $n$. 

---

1
It is said to be pointwise universal if for every stationary and ergodic source $X = (X_1, X_2, \ldots)$, with probability one,
\[
\frac{1}{n} l_n(X^n) \longrightarrow H(X).
\] (2)

As is common in the source coding literature (cf., e.g., [6, 2] and references therein), we use the term “pointwise” to indicate a sample path property. In the context of these notions of universality it suffices to restrict attention to stationary and ergodic sources since:

**Fact 1** If a sequence of schemes satisfies (1) for all stationary and ergodic sources then it is universal, i.e., it satisfies (1) for stationary and not necessarily ergodic sources as well.

The proof of Fact 1, which is little more than an interchange of order of integration, is given for completeness in Section 4.

The existence of pointwise universal codes is well known. One celebrated example is the Lempel-Ziv scheme [9]. Additional examples can be found in [7, 4] and references therein.

That pointwise universality of a code sequence implies its universality is a direct consequence of the bounded convergence theorem (and Fact 1). In that sense, pointwise universality seems like a stronger and more desirable property that plain universality. We are not aware, however, of lossless source codes in the literature that are universal but not pointwise universal. This raises the question of whether pointwise universality is a strictly stronger property.

In this correspondence we answer this question affirmatively by constructing a universal sequence of codes that is not pointwise universal.

Note that it is easy to find a code sequence and stationary ergodic sources for which (1) holds yet (2) does not. The issue is to find a code sequence satisfying (1) for all stationary ergodic sources yet failing to satisfy (2) on at least one such source.

## 2 The Idea

When randomized codes\(^3\) are allowed it is easy to construct a code sequence which is universal but not pointwise universal. Indeed, take any universal code sequence $\{C_n\}$ and define the sequence of randomized codes $\{\tilde{C}_n\}$ as follows: use $C_n$ with probability $1 - \varepsilon_n$ and send the source sequence uncompressed with probability $\varepsilon_n$ (plus an additional bit to convey mode of operation). Letting $\{\tilde{l}_n\}$ denote the associated length functions clearly
\[
\frac{1}{n} E\tilde{l}_n(X^n) = (1 - \varepsilon_n) \frac{1}{n} E l_n(X^n) + \varepsilon_n + \frac{1}{n}
\] (3)

so, as long as $\varepsilon_n \to 0$, the universality of $\{C_n\}$ implies that of $\{\tilde{C}_n\}$. On the other hand (assuming the randomization is performed independently for different values of $n$), if $\sum_n \varepsilon_n = \infty$ then, by the second Borel-Cantelli lemma (cf., e.g., [3]), for any source, with probability one,
\[
\limsup_{n \to \infty} \frac{1}{n} \tilde{l}_n(X^n) = 1.
\] (4)

Thus, not only is there positive probability that this randomized code sequence will not attain the entropy rate (assuming source with $H(X) < 1$), this event has probability one. Furthermore, (4) holds for all sources, whereas

\(^3\)Formally, a randomized lossless source code for $n$-tuples can be defined as a mapping $C_n : [0,1] \times \{0,1\}^n \to \{0,1\}^*$ with the property that $C_n(u, x^n) = C_n(v, y^n)$ for some $u, v \in [0,1]$ implies that $x^n = y^n$ ($u, v$ represent realizations of the randomization variable).
to negate pointwise universality it would have been enough to show that it holds for just one stationary and ergodic source.

Our idea in constructing a bona fide (non-randomized) universal but not pointwise universal code sequence is to use part of the source itself to effectively perform the kind of randomization artificially achieved by the scheme above. More specifically, suppose we are given a universal code sequence \( \{C_n\} \). For a sequence of mappings \( f_n : \{0, 1\}^n \to \{0, 1\} \) we can construct \( \{\tilde{C}_n\} \) as follows: use \( C_n \) if \( f_n(X^n) = 0 \) and send the source sequence uncompressed otherwise (plus an additional bit to convey mode of operation). The length functions satisfy a relationship similar to (3), where \( E_l(X^n) \) is replaced by \( E[l(X^n) \mid f_n(X^n) = 0] \) and \( \varepsilon_n = \Pr(f_n(X^n) = 1) \). Furthermore, similarly as for the randomized schemes, it follows from the second Borel-Cantelli lemma that if \( \{f_n\} \) can be chosen such that for some stationary ergodic source (of entropy rate less than 1) \( \sum \varepsilon_n = \infty \), and \( \{f_n(X^n)\} \) are mutually independent, then the resulting code sequence \( \{\tilde{C}_n\} \) is not pointwise universal. However, for it to remain universal, it must hold that \( \varepsilon_n \to 0 \) under all stationary sources.

A priori it is not clear that this approach can be fruitful since, conceivably, for \( \{f_n\} \) satisfying the first requirement one can find a source on which the second requirement fails. As is seen in our construction below, however, an \( \{f_n\} \) complying with both requirements exists.

### 3 Construction

Let \( \{C_n\} \) be a universal code sequence with associated length functions \( \{l_n\} \) and construct the bona fide (non-randomized) code sequence \( \{\tilde{C}_n\} \) as follows:

\[
\tilde{C}_n = C_n \quad \text{for} \quad \lfloor \log_2 n \rfloor \neq \log_2 n. \tag{5}
\]

When \( n = 2^k \) for an integer \( k \) let

\[
\tilde{C}_n(x^n) = \begin{cases} 0C_n(x^n) & \text{if} \ x^n = 0^n \text{ or } x_{m_k+1}^{m_k+1-1} \neq 0^m_{k+1-m_k} \\ 1x^n & \text{if} \ x^n \neq 0^n \text{ and } x_{m_k+1}^{m_k+1-1} = 0^m_{k+1-m_k}, \end{cases} \tag{6}
\]

where \( 0C_n(x^n) \) denotes \( 0 \) concatenated by \( C_n(x^n) \), \( 1x^n \) denotes \( 1 \) concatenated by \( x^n \), \( 0^l \) denotes the “all-zero” word of length \( l \), and \( \{m_k\}_k \) is constructed by letting \( m_0 = 0 \) and

\[
m_{k+1} = m_k + \max \left\{ \left\lfloor \frac{\log k}{\log 4} \right\rfloor + 1 \right\}. \tag{7}
\]

Note that \( \{m_k\}_k \) is strictly increasing, satisfies \( m_{k+1} - 1 \leq 2^k \) (so that \( x_{m_k+1}^{m_k+1-1} \) is a substring of \( x_2^k \)), and

\[
m_{k+1} - m_k \to \infty \quad \text{as} \quad k \to \infty. \tag{8}
\]

**Claim 1** The code sequence \( \{\tilde{C}_n\} \) is universal.

**Claim 2** The code sequence \( \{\tilde{C}_n\} \) is not pointwise universal.

### 4 Proofs

**Proof of Claim 1:** By Fact 1 it will suffice to establish \( E_{\frac{1}{n}l_n}(X^n) \to \overline{H}(X) \) assuming \( X \) stationary and ergodic. Letting below \( k = k_n = \lfloor \log_2 n \rfloor \),

\[
E_{\frac{1}{n}}l_n(X^n) \tag{9}
\]
Combining with (10) establishes

\[ E \left[ \frac{1}{n} \tilde{I}_n(X^n) \left| X^n = 0^n \text{ or } X_{m_k}^{m_k+1-1} \neq 0^{m_k+1-m_k} \right. \right] \Pr \left( X^n = 0^n \text{ or } X_{m_k}^{m_k+1-1} \neq 0^{m_k+1-m_k} \right) \\
+ E \left[ \frac{1}{n} \tilde{I}_n(X^n) \left| X^n \neq 0^n \text{ and } X_{m_k}^{m_k+1-1} = 0^{m_k+1-m_k} \right. \right] \Pr \left( X^n \neq 0^n \text{ and } X_{m_k}^{m_k+1-1} = 0^{m_k+1-m_k} \right) \\
\leq E \left[ \frac{1}{n} \tilde{I}_n(X^n) \left| X^n = 0^n \text{ or } X_{m_k}^{m_k+1-1} \neq 0^{m_k+1-m_k} \right. \right] + \frac{1}{n} \\
+ \frac{n+1}{n} \Pr \left( X^n \neq 0^n \text{ and } X_{m_k}^{m_k+1-1} = 0^{m_k+1-m_k} \right),
\]

where the inequality follows since, by construction, \( \frac{1}{n} \tilde{I}_n(X^n) \leq \frac{1}{n} I_n(X^n) + \frac{1}{n} \) whenever \( X^n = 0^n \) or \( X_{m_k}^{m_k+1-1} \neq 0^{m_k+1-m_k} \). Now we claim that for any stationary ergodic \( X \)

\[ \Pr \left( X^n \neq 0^n \text{ and } X_{m_k}^{m_k+1-1} = 0^{m_k+1-m_k} \right) \to 0 \text{ as } n \to \infty. \tag{11} \]

To see this consider one of two possible cases:

1. \( X \) is the “all-zero” source: In this case (11) trivially holds.

2. \( X \) is stationary ergodic and is not the “all-zero” source: In this case \( P(X_i = 1) > 0 \) so \( \Pr \left( X_i = 0^d \right) < \Pr \left( \frac{1}{n} \sum_{i=1}^{n} X_i \leq P(X_i = 1)/2 \right) \) implying, by ergodicity, \( \lim_{n \to \infty} \Pr \left( X^l = 0^d \right) = 0 \). It follows that (11) holds as well since it is even the case that \( \Pr \left( X_{m_k}^{m_k+1-1} = 0^{m_k+1-m_k} \right) \to 0 \) as \( n \to \infty \) (which, in turn, is a consequence of \( \lim_{n \to \infty} \Pr \left( X^l = 0^d \right) = 0 \) and the fact that \( m_{k+1} - m_k \to \infty \)).

Combining with (10) establishes

\[ \limsup_{n \to \infty} E \frac{1}{n} \tilde{I}_n(X^n) \leq \limsup_{n \to \infty} \left[ E \left[ \frac{1}{n} I_n(X^n) \left| X^n = 0^n \text{ or } X_{m_k}^{m_k+1-1} \neq 0^{m_k+1-m_k} \right. \right. \right]. \tag{12} \]

However, (11) also implies that

\[ \lim_{n \to \infty} \left( E \left[ \frac{1}{n} I_n(X^n) \left| X^n = 0^n \text{ or } X_{m_k}^{m_k+1-1} \neq 0^{m_k+1-m_k} \right. \right. \right] - E \frac{1}{n} l_n(X^n) \right) = 0 \tag{13} \]

which, in turn, combined with the universality of \( \{C_n\} \) implies

\[ \limsup_{n \to \infty} E \frac{1}{n} \tilde{I}_n(X^n) \leq \overline{H}(X) \tag{14} \]

for any \( X \) stationary and ergodic. Combined with the fact that \( \liminf_{n \to \infty} E \frac{1}{n} \tilde{I}_n(X^n) \geq \overline{H}(X) \) (which is the case for any code sequence), this completes the proof. \( \square \)

**Proof of Claim 2:** It will suffice to find one stationary ergodic source for which the asymptotic compression ratio using \( \{\tilde{C}_n\} \) does not converge with probability one to its entropy. To this end let \( X \) be the i.i.d. Bernoulli(3/4) source. Then, for every \( k \),

\[ \Pr \left( X_{m_k}^{m_k+1-1} = 0^{m_k+1-m_k} \right) = \frac{1}{4^{m_k+1-m_k}}. \tag{15} \]

Equation (7) implies that the right side of (15) is not summable (as a series in \( k \)), implying by the independence of the tuples \( \{X_{m_k}^{m_k+1-1}\}_k \) and the second Borel-Cantelli lemma that, with probability one,

\[ X_{m_k}^{m_k+1-1} = 0^{m_k+1-m_k} \text{ for infinitely many } k. \tag{16} \]

On the other hand,

\[ \Pr \left( X^{2^k} = 0^{2^k} \right) = \frac{1}{4^{2^k}}, \tag{17} \]
which is summable so by the first Borel-Cantelli lemma, with probability one,

$$X^{2^k} \neq 0^{2^k} \quad \text{for all sufficiently large } k.$$  \hspace{1cm} (18)

The combination of (16) and (18) implies that, with probability one,

$$X^{2^k} \neq 0^{2^k} \quad \text{and} \quad X^{m_{k+1}-1}_{m_k} = 0^{m_{k+1}-m_k} \quad \text{for infinitely many } k,$$

implying, by the construction of $\tilde{C}_n$ (recall (6)), that

$$\tilde{l}_n(X^n) = n + 1 \quad \text{infinitely often.} \hspace{1cm} (20)$$

Consequently, with probability one,

$$\limsup_{n \to \infty} \frac{1}{n} \tilde{l}_n(X^n) \geq 1,$$

which is strictly larger than the entropy rate of $X$. \hspace{1cm} \Box

Our only remaining debt is:

**Proof of Fact 1:** Let $\{L_n\}$ denote length functions of a code sequence satisfying

$$\lim_{n \to \infty} E_1 \frac{1}{n} L_n(X^n) = \overline{H}(X)$$

for all stationary and ergodic sources. Let $X \sim P$ be stationary, $\Theta$ denote the ergodic mode, and $\mu$ denote the distribution of $\Theta$ (so that $P = \int P_\theta d\mu(\theta)$ where $P_\theta$ is ergodic). Then (22) implies that for every $\theta$

$$\lim_{n \to \infty} E_\theta \frac{1}{n} L_n(X^n) = \overline{H}_\theta(X),$$

where $E_\theta$ and $\overline{H}_\theta$ denote, respectively, expectation and entropy rate under the source $P_\theta$. For every $\theta$ clearly $|\overline{H}_\theta(X)| \leq 1$ and $|E_\theta \frac{1}{n} L_n(X^n)| \leq c$ so, by bounded convergence,

$$\lim_{n \to \infty} E_\theta \frac{1}{n} L_n(X^n) = \int \overline{H}_\theta(X)d\mu(\theta).$$  \hspace{1cm} (24)

Now

$$\int \overline{H}_\theta(X)d\mu(\theta) = \int \left[ \lim_{n \to \infty} \frac{1}{n} H(X^n|\Theta = \theta) \right] d\mu(\theta)$$

$$= \int \left[ \inf_{n} \frac{1}{n} H(X^n|\Theta = \theta) \right] d\mu(\theta)$$

$$\leq \inf_{n} \frac{1}{n} \int H(X^n|\Theta = \theta) d\mu(\theta)$$

$$= \inf_{n} \frac{1}{n} H(X^n|\Theta)$$

$$\leq \inf_{n} \frac{1}{n} H(X^n)$$

$$= \overline{H}(X),$$  \hspace{1cm} (25)

where both (25) and (26) follow from the fact that for any stationary process $Y$, $\overline{H}(Y) = \inf_n \frac{1}{n} H(Y^n)$. Equations (26) and (24) imply $\lim_{n \to \infty} E_\theta \frac{1}{n} L_n(X^n) \leq \overline{H}(X)$ and complete the proof since the reverse equality holds for any code sequence. Note that this proof also implies $\int \overline{H}_\theta(X)d\mu(\theta) = \overline{H}(X).$  \hspace{1cm} \Box
5 Conclusion

We have shown that not all universal codes are pointwise universal.

The idea behind the construction extends to imply an analogous conclusion for other problems such as lossy coding, prediction, filtering, and denoising.

An interesting question is whether there exists an easily verified condition under which universality implies pointwise universality.

Acknowledgement

The author is grateful to George Gemelos for posing the question answered in this work.

References


