On the Entropy Rate of Pattern Processes

George M. Gemelos, Tsachy Weissman
Information Theory Research
HP Laboratories Palo Alto
HPL-2004-159
September 17, 2004*

entropy, pattern

This work examines some information theoretic properties of pattern sequences induced by stochastic processes. In particular, we study relationships between the entropy rate of the original process and that of its pattern sequence. We also characterize possible growth rates of the block entropy of pattern sequences.

On the Entropy Rate of Pattern Processes

George M. Gemelos Tsachy Weissman

September 17, 2004

Abstract

This work examines some information theoretic properties of pattern sequences induced by stochastic processes. In particular, we study relationships between the entropy rate of the original process and that of its pattern sequence. We also characterize possible growth rates of the block entropy of pattern sequences.

1 Introduction

In their recent work [1], Orlitsky et. al discuss the compression of sequences with unknown alphabet size. This work, among others, has created interest in examining random processes with arbitrary alphabets which may a priori be unknown. One can think of this as a problem of reading a foreign language for the first time. As one begins to parse characters, one's knowledge of the alphabet grows. Since the characters in the alphabet have initially no meaning beyond the order in which they appear, one can relabel these characters by the order of their first appearance. Given a string, we refer to the relabelled string as the *pattern* associated with the original string.

Example 1 *Assume that the following English sentence was being parsed into a pattern by a non-English speaker.*

english is hard to learn...

The associated pattern would be

$$1, 2, 3, 4, 5, 6, 7, 8, 5, 6, 8, 7, 9, 10, 11, 8, 12, 13, 8, 4, 1, 9, 10, 2, \ldots$$

regarding the space too as a character.

We abstract this as follows: Given a stochastic process $\{X_i\}_{i\geq 1}$, we create a pattern process $\{Z_i\}_{i\geq 1}$.

It is the compression of the pattern process $\{Z_i\}$ that Orlitsky et. al. focus on in [1]. They justify this emphasis by reasoning that the bulk of the information is in the pattern. Although universal compression is an extensively studied problem, the universal compression of pattern sequences is relatively new, see [1], [2], [3], [4], [5] and [6]. These recent papers address universality questions of how well a pattern sequence associated with an unknown source can be compressed relative to the case where this distribution is known. Emphasis is on quantifying the redundancy, i.e., the difference between what can be achieved without and with knowledge of the source distribution. In this work we restrict our attention to the second term of the said difference, namely, to the entropy of the pattern sequence. More specifically, our goal is to study some of the relationships between the entropy rate $H(\mathbf{X})$ of the original process[1] $\{X_i\}$, and the entropy rate $H(\mathbf{Z})$ of the associated pattern process. This relationship is not always trivial, as the following examples illustrate.

Example 2 *Let X_i be drawn i.i.d. $\sim P$, where P is a pmf on a finite alphabet. Then we show below that $H(\mathbf{X}) = H(\mathbf{Z})$.*

The intuition behind this result is that given enough time, all the symbols with positive probability will be seen, after which time the original process and its associated pattern sequence coincide, up to relabelling of the alphabet symbols.

Example 3 *Let X_i be drawn i.i.d. $\sim$ uniform $[0, 2]$. Then $H(\mathbf{X}) = \infty$. Since the probability of seeing the same number twice is zero, $Z_i = i$ with probability 1 for all i and, consequently, $H(\mathbf{Z}) = 0$.*

The above two are extreme examples illustrating the fact that the relationship between $H(\mathbf{X})$ and $H(\mathbf{Z})$ is not trivial. In Section 2 we characterize this relationship for the case of a generally distributed *i.i.d.* process, as well as in various other cases involving Markov processes, stationary ergodic processes, and hidden Markov processes (under certain restrictions on the associated source alphabets). In Section 3 we characterize a set of achievable asymptotic growth rates for the block entropy of a pattern process. We conclude in Section 4 with a brief summary of our results and a conjecture.

2 Entropy Rates

Consider first the case where $\{X_i\}$ are generated *i.i.d.* $\sim f$, where f is an arbitrary distribution on the arbitrary source alphabet $\mathcal{X}$. In general, f can be decomposed into two parts: a part consisting solely of point masses, and one consisting of a distribution that does not

[1] $H(\mathbf{X})$ will denote entropy rate throughout this work, regardless of the discreteness of the distributions of $\{X^n\}$. It is thus to be regarded as ∞ when these are not discrete.

contain any point masses. Let S_f be the set of all point masses $S_f = \{x \in \mathcal{X} : f(x) > 0\}$. There exists a pmf f_m, on S_f, and a distribution with no point masses f_d, such that

$$f = \alpha f_m + \bar{\alpha} f_d,$$

where $\alpha = f(S_f)$.

Theorem 1 *For $\{X_i\}$ i.i.d. $\sim f$, and any point $x_0 \in S_f^c$ let*

$$f^* = \alpha f_m + \bar{\alpha}\delta_{x_0},$$

where δ_{x_0} denotes the probability distribution assigning probability one to x_0. Then

$$H(\mathbf{Z}) = H(f^*),$$

where $H(f^*)$ denotes the entropy of the discrete distribution f^*.

The proof of Theorem 1 is deferred to the Appendix (as it will employ a corollary more naturally proved in the next section). As can be seen, Theorem 1 is consistent with Example 2 and Example 3. Note that f^* is created by taking all the point masses in f and assigning all the remaining probability to a new point mass. This corresponds to the result in Example 3 which suggests that the pattern of a process drawn according to a pdf has no randomness, i.e. an entropy rate of zero. Therefore, the only randomness in the pattern comes from the point masses and the event of falling on a "non-point-mass-mode".

Example 4 *Let $\{X_i\}$ be i.i.d. with each component drawn, with probability $1/3$, as a $N(0,1)$ and, with probability $2/3$, as a $Bern(1/2)$. Applying Theorem 1 for this case gives that f^* is the uniform distribution on an alphabet of size 3. Therefore, $H(\mathbf{Z}) = \log(3)$.*

Although the number of point masses in the f^*-s associated with all three examples above are finite, it is important to note that Theorem 1 makes no such assumption.

Our natural next step is the case where $\{X_i\}$ is generated by a Markov process. The entropy rate of Markov processes is well-known. What can be said about the entropy rate of the associated pattern process? We can begin in this context with an even more general setting, but for the case of a finite alphabet.

Theorem 2 *Let $\{X_i\}$ be a stationary ergodic process with components in the alphabet $\mathcal{X}$, where $|\mathcal{X}| < \infty$, and let $\{Z_i\}$ be the associated pattern process. Then $H(\mathbf{X}) = H(\mathbf{Z})$.*

Proof: Define the mapping $g^n : \mathcal{X} \mapsto \{1, \ldots, |\mathcal{X}|\} \cup \infty$ where

$$g^n(x) = \inf\{Z_i : i \leq n, X_i = x\},$$

where an infimum over an empty set is defined as ∞. We can think of g^n as the label of x when it appears in the sequence X^n.

$$\begin{aligned}
H(\mathbf{Z}) &= \lim_{n\to\infty} \frac{H(Z^n)}{n} \\
&\geq \lim_{n\to\infty} \frac{H(Z^n|g^n)}{n} \\
&\stackrel{(a)}{=} \lim_{n\to\infty} \frac{H(X^n|g^n)}{n} \\
&= \lim_{n\to\infty} \frac{H(X^n, g^n)}{n} - \lim_{n\to\infty} \frac{H(g^n)}{n} \\
&\geq \lim_{n\to\infty} \frac{H(X^n)}{n} - \lim_{n\to\infty} \frac{H(g^n)}{n} \\
&\stackrel{(b)}{=} \lim_{n\to\infty} \frac{H(X^n)}{n} - \lim_{n\to\infty} \frac{|\mathcal{X}|\log(|\mathcal{X}|+1)}{n},
\end{aligned}$$

where (a) comes from the fact that $\{X^n\}$ is a deterministic function of $\{Z^n\}$ given g^n and (b) from the fact that there are at most $|\mathcal{X}|^{|\mathcal{X}|}$ possible maps g^n. Thus we got

$$H(\mathbf{Z}) \geq H(\mathbf{X}). \tag{1}$$

The upper bound $H(X^n) \geq H(Z^n)$ holds for all n from the data processing inequality and, hence, $H(\mathbf{X}) \geq H(\mathbf{Z})$. Combining this with (1) gives $H(\mathbf{X}) = H(\mathbf{Z})$. □

We now look at the case of a first order Markov process with components in a countable alphabet.

Theorem 3 *Let $\{X_i\}$ be a stationary ergodic first order Markov process on the countable alphabet $\mathcal{X}$ and let $\{Z_i\}$ be the associated pattern process. Then $H(\mathbf{X}) = H(\mathbf{Z})$.*

Proof: Let μ be the stationary distribution of the Markov process and let $P_x(y) = P(X_{t+1} = y|X_t = x)$ for all $x, y \in \mathcal{X}$. The data processing inequality implies $H(X^n) \geq H(Z^n)$ for all n. Hence $H(\mathbf{X}) \geq H(\mathbf{Z})$. To complete the proof it remains to show $H(\mathbf{X}) \leq H(\mathbf{Z})$, for which we will need the following elementary fact and the lemma following it.

Fact 1 *Let $\{A_n\}$ and $\{B_n\}$ be two sequences of events such that $\lim_{n\to\infty} P(A_n) = 1$ and $\lim_{n\to\infty} P(B_n) = b$. Then $\lim_{n\to\infty} P(A_n \cap B_n) = b$.*

For completeness we provide a proof of this elementary fact in the Appendix.

Lemma 1 *Given any $B \subseteq \mathcal{X}$ such that $|B| < \infty$*

$$H(\mathbf{Z}) \geq \sum_{b\in B} \mu(b) H(\Phi_B(P_b)),$$

where $\Phi_B(P_x) \triangleq \mu(B)P_x + \mu(B^c)\delta_{x_0}$, for an arbitrary $x_0 \notin B$.

Proof of lemma: Let $A(x^n) \triangleq \{x_1, \ldots, x_n\}$.

$$\begin{aligned}
H(\mathbf{Z}) &= \lim_{n\to\infty} H(Z_n|Z^{n-1}) \\
&\geq \lim_{n\to\infty} H(Z_n|X^{n-1}) \\
&\geq \lim_{n\to\infty} \sum_{b\in B} \int_{\{x^{n-1}: B\subseteq A(x^{n-1}),\ x_{n-1}=b\}} H(P(Z_n|X^{n-1}=x^{n-1}))\ dP_{X^{n-1}} \\
&\geq \lim_{n\to\infty} \sum_{b\in B} P(B\subseteq A(X^{n-1}), X_{n-1}=b) H(\Phi_B(P_b)) \\
&= \sum_{b\in B} H(\Phi_B(P_b)) \lim_{n\to\infty} P(B\subseteq A(X^{n-1}), X_{n-1}=b) \\
&\overset{(a)}{\geq} \sum_{b\in B} \mu(b) H(\Phi_B(P_b))
\end{aligned}$$

where (a) is a consequence of Fact 1. □

Let now $\{B_k\}$ be a sequence of sets such that $B_k \subseteq \mathcal{X}$, $|B_k| < \infty$ for all k, and

$$\lim_{k\to\infty} \sum_{b\in B_k} \sum_{a\in B_k} -\mu(b) P_b(a) \log P_b(a) = \sum_{b\in\mathcal{X}} \sum_{a\in\mathcal{X}} -\mu(b) P_b(a) \log P_b(a),$$

regardless of the finiteness of both sides of the equation. Note that since the above summands are all positive, such a sequence $\{B_k\}$ can always be found. Lemma 1 gives us

$$H(\mathbf{Z}) \geq \sum_{b\in B_k} \mu(b) H(\Phi_B(P_b)) \quad \forall\ k.$$

Hence, by taking $k \to \infty$, we get

$$\begin{aligned}
H(\mathbf{Z}) &\geq \lim_{k\to\infty} \sum_{b\in B_k} \mu(b) H(\Phi_B(P_b)) \\
&\geq \lim_{k\to\infty} \sum_{b\in B_k} \mu(b) \sum_{a\in B_k} -P_b(a) \log P_b(a) \\
&= \lim_{k\to\infty} \sum_{b\in B_k} \sum_{a\in B_k} -\mu(b) P_b(a) \log P_b(a) \\
&\overset{(a)}{=} \sum_{b\in\mathcal{X}} \sum_{a\in\mathcal{X}} -\mu(b) P_b(a) \log P_b(a) \\
&\overset{(b)}{=} H(\mathbf{X}),
\end{aligned} \tag{2}$$

where (a) comes from the construction of $\{B_k\}$ and (b) from the fact that $\{X_i\}$ is a stationary first-order Markov process. □

One should note that the proof of Theorem 3 can easily be extended to the case of Markov processes of any order. Hence, without going through the proof, we state the following:

Theorem 4 *Let $\{X_i\}$ be a stationary ergodic Markov process (of any order) on the countable alphabet $\mathcal{X}$, and let $\{Z_i\}$ be the associated pattern process. Then $H(\mathbf{X}) = H(\mathbf{Z})$.*

We now consider the case of a noise-corrupted process. Let X_i be a stationary ergodic process and Y_i be its noise corrupted version. Here we assume *i.i.d.* additive noise, N_i. We will also assume that X_i takes values in a finite alphabet $\mathcal{A} \subseteq \mathbb{R}$. Let S_X, S_Y and S_N denote the set of points of positive measure for X_i, Y_i and N_i respectively (assumed to all take values in $\mathbb{R}$). We will also define the discrete random variable

$$\tilde{N}_i = N_i 1_{\{N_i \in S_N\}} + n_o 1_{\{N_i \in S_N^c\}},$$

for an arbitrary point $n_0 \notin S_N$.

Theorem 5 *Let $\{X_i\}$ be a stationary ergodic process. Let $\{Y_i\}$ and $\{\tilde{Y}_i\}$ denote, respectively, the process $\{X_i\}$ corrupted by the additive noise $\{N_i\}$ and $\{\tilde{N}_i\}$. If $|S_Y| < \infty$, then the entropy rate of the pattern processes associated, respectively, with $\{Y_i\}$ and $\{\tilde{Y}_i\}$, $\{Z_i\}$ and $\{\tilde{Z}_i\}$, are equal.*

Proof: Define

$$\hat{Z}(n)_i = Z_i 1_{\{\exists j \in [1,n] \setminus i:\ Z_i = Z_j\}} + y_o 1_{\{\exists j \in [1,n] \setminus i:\ Z_i = Z_j\}^c}$$

for some arbitrary $y_0 \notin S_Y$. Clearly $\hat{Z}(n)_i$ uniquely determines Z^n and vice versa so, in particular,

$$H(Z^n) = H(\hat{Z}(n)) \quad \forall\ n > 0. \tag{3}$$

We also observe that we can construct $\hat{Z}(n)$ from $\tilde{Z}^n$ w.p. 1. Therefore

$$H(\tilde{Z}^n) \geq H(\hat{Z}(n)) \quad \forall\ n > 0. \tag{4}$$

Combining (3) and (4) gives

$$H(\tilde{\mathbf{Z}}) \geq H(\mathbf{Z}). \tag{5}$$

Defining $C(n)_i = 1_{\{\hat{Z}(n)_i = y_0\} \cap \{Y_i \in S_Y\}}$ we observe that given $C(n)$ and $\hat{Z}(n)$ we can reconstruct $\tilde{Z}^n$ for all $n > 0$. Hence, for all $n > 0$,

$$\begin{aligned} H(\tilde{Z}^n) &\leq H(\hat{Z}(n), C(n)) \\ &\leq H(\hat{Z}(n)) + H(C(n)) \\ &\stackrel{(a)}{=} H(Z^n) + H(C(n)) \\ &\leq H(Z^n) + \sum_{i=1}^{n} H(C(n)_i), \end{aligned} \tag{6}$$

where (a) comes from (3).

Let

$$Pe_i^{(n)} = Pr\{Y_i \in S_Y,\ Y_j \neq Y_i \ \forall\ j \in [i,n]\backslash i\}.$$

Then we have

$$\begin{aligned} Pe_i^{(n)} &= Pr\{Y_i \in S_Y\} Pr\{Y_j \neq Y_i \ \forall\ j \in [1,n]\backslash i \ |Y_i \in S_Y\} \\ Pe_i^{(n)} &\overset{(a)}{=} Pr\{S_Y\} \sum_{y \in S_Y} Pr\{Y_j \neq y \ \forall\ j \in [1,n]\backslash i \ |Y_i = y\}, \end{aligned}$$

where (a) comes from the stationarity of Y. Without loss of generality assume that $i > n/2$

$$\begin{aligned} Pe_i^{(n)} &\leq Pr\{S_Y\} \sum_{y \in S_Y} Pr\{Y_j \neq y \ \forall\ j \in [i - n/2 + 1, i - 1]|Y_i = y\} \\ Pe_i^{(n)} &\overset{(b)}{\leq} Pr\{S_Y\} \sum_{y \in S_Y} Pr\{Y_j \neq y \ \forall\ j \in [2, n/2]|Y_1 = y\}, \end{aligned}$$

where (b) comes from the stationarity of Y. Let

$$Pe^{(n)} = Pr\{S_Y\} \sum_{y \in S_Y} Pr\{Y_j \neq y \ \forall\ j \in [2, n/2]|Y_1 = y\}. \tag{7}$$

Therefor we have

$$Pe^{(n)} \geq Pe_i^{(n)} \quad \forall i. \tag{8}$$

Since $|S_Y| < \infty$, by ergodicity we have

$$\lim_{n\to\infty} Pr\{Y_j \neq y \ \forall\ j \in [2, n/2]|Y_1 = y\} = 0$$

and (7) gives us $\lim_{n\to\infty} Pe^{(n)} = 0$. Hence there exists an N such that $Pe^{(n)} < 1/2$ for all $n > N$ and (8) implies that

$$H_B(Pe_i^{(n)}) \leq H_B(Pe^{(n)}) \quad \forall n \geq N, \tag{9}$$

where H_B is the binary entropy function. Substituting $Pe_i^{(n)}$ into (6) and taking the normalized limits we get

$$\lim_{n\to\infty} \frac{H(\tilde{Z}^n)}{n} \leq \lim_{n\to\infty} \frac{H(Z^n)}{n} + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H_B(Pe_i^{(n)})$$

and, (9) gives us

$$\begin{aligned} \lim_{n\to\infty} \frac{H(\tilde{Z}^n)}{n} &\leq \lim_{n\to\infty} \frac{H(Z^n)}{n} + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H_B(Pe^{(n)}) \\ &\leq \lim_{n\to\infty} \frac{H(Z^n)}{n} + \lim_{n\to\infty} H_B(Pe^{(n)}) \end{aligned}$$

and, since $\lim_{n\to\infty} Pe^{(n)} = 0$,

$$\lim_{n\to\infty} \frac{H(\tilde{Z}^n)}{n} \le \lim_{n\to\infty} \frac{H(Z^n)}{n} + 0.$$

Therefore

$$H(\tilde{\mathbf{Z}}) \le H(\mathbf{Z}). \tag{10}$$

Combining (5) and (10) completes the proof. □

The following is directly implied by Theorems 2 and 5.

Corollary 6 *Let X_i be a stationary ergodic process,*

$$Y_i = X_i + N_i$$

where N_i is an i.i.d. sequence and S_Y and $\tilde{Y}_i$ are defined as in Theorem 5. Assume further $|S_Y| < \infty$. If X_i takes values in a finite alphabet then

$$H(\mathbf{Z}) = H(\tilde{\mathbf{Y}}).$$

3 Growth Rates

We now turn our attention to the asymptotic growth rate for the block entropy of a pattern sequence. We begin by stating our main result which is a set of such achievable growth rates.

Proposition 1 *For any $\delta > 0$ there exists a process $\{X_i\}$ such that its associated pattern sequence satisfies*

$$\lim_{n\to\infty} \frac{H(Z_{n+1}|Z^n)}{(\ln n)^{1-\delta}} = \infty. \tag{11}$$

Before we begin the proof of Proposition 1 we need to prove some useful facts. Let X_i be *i.i.d.* $\sim X$, where X takes values in an arbitrary space $\mathcal{X}$, and $\{Z_i\}_{i\ge 1}$ be the associated pattern sequence. Define $\mathcal{D} = \{x \in \mathcal{X} : \Pr(X = x) > 0\}$. For $\mathcal{B} \subseteq \mathcal{D}$ let $P^{\mathcal{B}}$ denote the (point-mass) distribution on $\mathcal{B} \cup \{s\}$ (where s stands for "special symbol") with

$$P^{\mathcal{B}}(x) = \begin{cases} \Pr(X = x) & \text{for } x \in \mathcal{B} \\ 1 - \sum_{x\in\mathcal{B}} \Pr(X = x) & \text{for } x = s \end{cases}$$

and let

$$h(P^{\mathcal{B}}) = \sum_{x\in\mathcal{B}\cup\{s\}} P^{\mathcal{B}}(x) \log \frac{1}{P^{\mathcal{B}}(x)}$$

denote its entropy.

Claim 1 *$h(P^{\mathcal{B}})$ is increasing in $\mathcal{B}$, i.e., for any $\mathcal{B}_1 \subseteq \mathcal{B}_2 \subseteq \mathcal{D}$*

$$h(P^{\mathcal{B}_1}) \leq h(P^{\mathcal{B}_2}).$$

Proof: This is nothing but a data-processing inequality. Indeed, let $Y \sim P^{\mathcal{B}_2}$ and let

$$U = \begin{cases} Y & \text{if } Y \in \mathcal{B}_1 \\ s & \text{otherwise.} \end{cases}$$

Clearly $U \sim P^{\mathcal{B}_1}$ and U is a deterministic function of Y, thus the claim follows. □

Define now further, for $x^n \in \mathcal{X}^n$,

$$\mathcal{A}(x^n) = \{a \in \mathcal{D} : x_i = a \text{ for some } 1 \leq i \leq n\}.$$

In words, $\mathcal{A}(x^n)$ are the elements of $\mathcal{D}$ that appear in x^n.

Proposition 2 *For any $\mathcal{B} \subseteq \mathcal{D}$*

$$H(Z_{n+1}|Z^n) \geq h\left(P^{\mathcal{B}}\right)\left[1 - |\mathcal{B}| \exp\left(-n \min_{b\in\mathcal{B}} \Pr(X = b)\right)\right].$$

Proof: Letting P_X^n denote the distribution of X^n, for any $\mathcal{B} \subseteq \mathcal{D}$,

$$\begin{aligned}
H(Z_{n+1}|Z^n) &\geq H(Z_{n+1}|X^n) && (12)\\
&= \int_{\mathcal{X}^n} H(Z_{n+1}|X^n = x^n) dP_X^n(x^n) && (13)\\
&= \int_{\mathcal{X}^n} h\left(P^{\mathcal{A}(x^n)}\right) dP_X^n(x^n) && (14)\\
&\geq \int_{\{x^n:\mathcal{B}\subseteq\mathcal{A}(x^n)\}} h\left(P^{\mathcal{A}(x^n)}\right) dP_X^n(x^n) && (15)\\
&\geq h\left(P^{\mathcal{B}}\right) \Pr\left(\mathcal{B} \subseteq \mathcal{A}(X^n)\right), && (16)
\end{aligned}$$

where the last inequality follows from the monotonicity property in Claim 1. Now, for any $\mathcal{B} \subseteq \mathcal{D}$,

$$\begin{aligned}
\Pr\left(\mathcal{B} \not\subseteq \mathcal{A}(X^n)\right) &= \Pr\left(\bigcup_{b\in\mathcal{B}} \{b \notin \mathcal{A}(X^n)\}\right) && (17)\\
&\leq \sum_{b\in\mathcal{B}} \Pr\left(b \notin \mathcal{A}(X^n)\right) && (18)\\
&= \sum_{b\in\mathcal{B}} (1 - \Pr(X = b))^n && (19)\\
&\leq |\mathcal{B}| \left(1 - \min_{b\in\mathcal{B}} \Pr(X = b)\right)^n && (20)\\
&\leq |\mathcal{B}| \exp\left(-n \min_{b\in\mathcal{B}} \Pr(X = b)\right). && (21)
\end{aligned}$$

The proposition now follows by combining (16) with (21). □

Corollary 7

$$\liminf_{n\to\infty} H(Z_{n+1}|Z^n) \geq h\left(P^{\mathcal{D}}\right),$$

regardless of the finiteness of the right side of the inequality.

Proof: Take a sequence $\{\mathcal{B}_k\}$ of finite subsets $\mathcal{B}_k \subseteq \mathcal{D}$ satisfying

$$\lim_{k\to\infty} h\left(P^{\mathcal{B}_k}\right) = h\left(P^{\mathcal{D}}\right).$$

Proposition 2 implies, for each k,

$$\liminf_{n\to\infty} H(Z_{n+1}|Z^n) \geq h\left(P^{\mathcal{B}_k}\right), \tag{22}$$

completing the proof by taking $k \to \infty$ on the right side of (22). □

Sketch of the proof of Proposition 1: Consider the case where $\{X_i\}$ are generate *i.i.d.* $\sim P$, where P is a distribution on $\mathbb{N}$ and $p_j = \Pr(X_i = j)$ is a non-increasing sequence. Letting $S_l = \sum_{i=1}^{l} p_i \log \frac{1}{p_i}$ it follows by taking $\mathcal{B} = \mathcal{B}_l = \{1, \ldots, l\}$ in Proposition 2 that

$$\begin{aligned} H(Z_{n+1}|Z^n) &\geq h\left(P^{\mathcal{B}_l}\right)\left[1 - |\mathcal{B}_l| \exp\left(-n \min_{b\in\mathcal{B}_l} \Pr(X = b)\right)\right] \\ &\geq S_l \left[1 - l \exp\left(-np_l\right)\right] \end{aligned}$$

implying, by the arbitrariness of l,

$$H(Z_{n+1}|Z^n) \geq \max_l S_l \left[1 - l \exp\left(-np_l\right)\right]. \tag{23}$$

Consider now the distribution

$$p_i = \Pr(X = i) = \frac{c(\varepsilon)}{i(\ln i)^{1+\varepsilon}}, \tag{24}$$

for some $\varepsilon > 0$, where $c(\varepsilon)$ is the normalization constant. In this case $S_l = \sum_{i=1}^{l} \frac{c(\varepsilon)}{i(\ln i)^{1+\varepsilon}} \log \frac{i(\ln i)^{1+\varepsilon}}{c(\varepsilon)} \gtrsim \sum_{i=1}^{l} \frac{1}{i(\ln i)^{\varepsilon}} \sim (\ln l)^{1-\varepsilon}$ thus (23) implies, taking $l \approx n^{(1-\varepsilon)/(1+\varepsilon)}$,

$$\begin{aligned} H(Z_{n+1}|Z^n) &\geq S_l \left[1 - l \exp\left(-np_l\right)\right] \\ &\gtrsim (\ln l)^{1-\varepsilon} \left[1 - l \exp\left(-np_l\right)\right] \\ &\gtrsim (\ln l)^{1-\varepsilon} \left[1 - l \exp\left(-n/l^{1+\varepsilon}\right)\right] \\ &\gtrsim (\ln n)^{1-\varepsilon} \left[1 - n \exp\left(-n^{\varepsilon}\right)\right] \\ &\sim (\ln n)^{1-\varepsilon}. \end{aligned}$$

Thus (11) is satisfied under the distribution in (24) with any $\varepsilon < \delta$. □

4 Conclusion

We have characterized relationships between the entropy rate of a source and that of its pattern process for the *i.i.d.* case, the case of a stationary ergodic finite-alphabet source, Markov processes of any order with countable state spaces, and additive noise sequences.

We also examined possible asymptotic growth rates for the block-entropy of pattern sequences. The following is inspired by Proposition 1.

Conjecture 1 *Let $\{X_i\}$ be an arbitrarily distributed stationary and ergodic process and let $\{Z_i\}$ be its pattern process. Then*

$$\lim_{n\to\infty} \frac{H(Z_{n+1}|Z^n)}{\log n} = 0.$$

On the other hand, for any $f(n) = o(\log n)$ there exists an i.i.d. process $\{X_i\}$ such that

$$\lim_{n\to\infty} \frac{H(Z_{n+1}|Z^n)}{f(n)} = \infty.$$

Regarding the first assertion in the conjecture note that the pattern sequence associated with an arbitrarily distributed source satisfies

$$\limsup_{n\to\infty} \frac{H(Z_{n+1}|Z^n)}{\log n} < \infty$$

since obviously $H(Z_{n+1}|Z^n) \leq \log(n+1)$ for all n. This part of the conjecture, then, asserts slightly more for the case where $\{X_i\}$ is stationary and ergodic, namely that in this case necessarily $H(Z_{n+1}|Z^n) = o(\log n)$. The second part of the conjecture, we believe, might be provable by refining the argument used in the proof of Proposition 1.

A Proof of Theorem 1

If $\alpha = 0$, then X_i is *i.i.d.* $\sim f_d$. Therefore $H(Z^n) = 0$ for all n. This gives us that $H(\mathbf{Z}) = 0$ which agrees with Theorem 1. Hence we just need to prove Theorem 1 for the case where $\alpha > 0$. Therefore, S_f exists and $|S_f| > 0$.

Note that Corollary 7 gives us $H(\mathbf{Z}) \geq H(f^*)$. For the reverse inequality, define the process $\{\hat{X}_i\}$ where

$$\hat{X}_i = X_i 1_{\{X_i \in S_f\}} + x_0 1_{\{X_i \notin S_f\}},$$

for some $x_0 \notin S_f$. Clearly, $\hat{X}_i$ is *i.i.d.* $\sim f^*$. We thus have

$$\begin{aligned} H(\mathbf{Z}) &= \lim_{n\to\infty} \frac{H(Z^n)}{n} \\ &\overset{(a)}{\leq} \lim_{n\to\infty} \frac{H(\hat{X}^n)}{n} \\ &\overset{(b)}{=} H(f^*), \end{aligned} \tag{25}$$

where (a) comes from the fact that Z^n is a deterministic function of $\hat{X}^n$ with probability 1, and (b) from the fact that $\hat{X}_i$ is $i.i.d. \sim f^*$. $\square$

Proof of Fact 1: Trivially, $P(A_n \cap B_n) \leq P(B_n) \to b$. On the other hand,

$$\begin{aligned}
\liminf_{n\to\infty} P(A_n \cap B_n) &= \liminf_{n\to\infty} 1 - P(A_n^c \cup B_n^c) \\
&\geq \liminf_{n\to\infty} 1 - P(A_n^c) - P(B_n^c) \\
&= 1 - 0 - (1-b).
\end{aligned}$$

$\square$

References

[1] A. Orlitsky, N.P. Santhanam, and J. Zhang, Universal compression of memoryless sources over unknown alphabets, IEEE Transactions on Information Theory, IT-50:7 (July 2004), pp. 1469-1481.

[2] N. Jevtić, A. Orlitsky, N. Santhanam, Universal compression of unknown alphabets, Proceedings of the 2002 IEEE International Symposium on Information Theory, p. 302.

[3] A. Orlitsky, N. P. Santhanam, Performance of universal codes over infinite alphabets, Proceedings of the 2003 Data Compression Conference.

[4] A. Orlitsky, N. P. Santhanam, J. Zhang, Bounds on compression of unknown alphabets, Proceedings of the 2003 IEEE International Symposium on Information Theory.

[5] G. I. Shamir, On the MDL Principle for Universal Compression of Unknown Alphabets, Proceedings of the 2002 Allerton Conference on Communication, Control, and Computing, pp. 1238-1247.

[6] G. I. Shamir and L. Song, On the entropy of patterns of I.I.D. sequences, Proceedings of the 2003 Allerton Conference on Communication, Control, and Computing