ELECTRONIC RESEARCH ANNOUNCEMENTS
OF THE AMERICAN MATHEMATICAL SOCIETY
Volume 3, Pages 11–16 (March 12, 1997)
S 1079-6762(97)00018-8

ERGODIC BEHAVIOR OF GRAPH ENTROPY

JOHN KIEFFER AND EN-HUI YANG

(Communicated by Douglas Lind)

ABSTRACT. For a positive integer n, let X^n be the vector formed by the first n samples of a stationary ergodic finite alphabet process. The vector X^n is hierarchically represented via a finite rooted acyclic directed graph G_n. Each terminal vertex of G_n carries a label from the process alphabet, and X^n can be reconstituted as the sequence of labels at the ends of the paths from root vertex to terminal vertex in G_n. The entropy $H(G_n)$ of the graph G_n is defined as a nonnegative real number computed in terms of the number of incident edges to each vertex of G_n. An algorithm is given which assigns to G_n a binary codeword from which G_n can be reconstructed, such that the length of the codeword is approximately equal to $H(G_n)$. It is shown that if the number of edges of G_n is $o(n)$, then the sequence $\{H(G_n)/n\}$ converges almost surely to the entropy of the process.

1. Introduction

In the hierarchical approach to data compression developed by the authors [2], [3], [4], finite rooted acyclic directed graphs can be used to represent the data strings that are to be compressed. To see how this representation works, let us determine the data string x represented by the graph in Figure 1. This graph contains ten edges labelled 1 through 10 and two terminal vertices labelled 0 and 1. If we list the paths in this graph that go from root vertex to a terminal vertex, we obtain the ten paths

$$(1),(2,5),(2,6),(3,7,5),(3,7,6),(3,8),(4,9,7,5),(4,9,7,6),(4,9,8),(4,10).$$

(The paths are listed in lexicographical order.) The string x is then obtained by replacing each path in this list with the label on the terminal vertex for that path. We see that $x = 0010110111$.

Let $G = (V, E)$ denote an arbitrary finite rooted acyclic directed graph, where V is the set of vertices and E is the set of edges. Let $|V|, |E|$ denote the number of vertices and the number of edges, respectively. (In general, let $|S|$ denote the cardinality of any finite set S.) For each $v \in V$, let $i(v)$ be the number of edges

Received by the editors December 12, 1996.
1991 *Mathematics Subject Classification.* Primary 28D99; Secondary 60G10, 94A15.
Key words and phrases. Graphs, entropy, compression, stationary ergodic process.
This work was supported in part by the National Science Foundation under Grants NCR-9304984 and NCR-9627965.

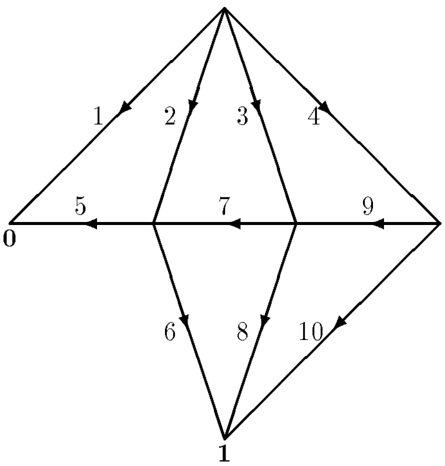

FIGURE 1

that terminate at v. Following [2], the entropy $H(G)$ of the graph G is defined by

$$H(G) = \sum_{v \in V,\, i(v) \geq 2} (i(v) - 1) \log \left(\frac{|E| - |V| + 1}{i(v) - 1} \right),$$

where the logarithm throughout shall be to base two. The entropy of the graph in Figure 1 is $3 \log 5 + 2 \log(5/2) = 9.61$.

Suppose we are given an infinite finite-alphabet sequence $(x_1, x_2, \dots)$, and let $x^n = (x_1, x_2, \dots, x_n), n \geq 1$. For each n, a graph G_n is selected to represent the data string x^n. One then compresses x^n by assigning a binary codeword to G_n that allows one to reconstruct G_n and therefore x^n. The length of this codeword is approximately equal to $H(G_n)$. (See Lemma 1.) A natural question is the behavior of the entropies $\{H(G_n)\}$ as $n \to \infty$ when the sequence $(x_1, x_2, \dots)$ is generated by a stationary ergodic process. In this case, our main result (Theorem 1) states that $\{H(G_n)/n\}$ converges almost surely to the entropy of the process, provided only that the number of edges of G_n is $o(n)$. Some of the many applications of this result shall be discussed.

2. MAIN RESULT

If B is a finite set, let B^* denote the set of all strings of finite length formed from symbols in B. Suppose $s_1, s_2, \dots, s_n$ are strings in a set B^*. Let $s_1 * s_2 * \cdots * s_n$ denote the string in B^* obtained by concatenating together the strings $s_1, s_2, \dots, s_n$ in the indicated order.

Let $G = (V, E)$ be a finite rooted acyclic directed graph. Let V_t denote the set of terminal vertices of G. For each $v \in V$, $v \notin V_t$, let $E^+(v)$ denote the set of all edges emanating from v. For each $v \in V$, $v \neq$ root vertex, let $E^-(v)$ denote the set of all edges which terminate at v. We say that G is a *canonical* graph if

(i): $V = \{1, 2, \dots, |V|\}$ and $1 \in V$ is the root vertex.
(ii): $E = \{1, 2, \dots, |E|\}$.
(iii): If $v_1, v_2 \in \{2, \dots, |V|\}$ and $v_1 < v_2$, then $\min E^-(v_1) < \min E^-(v_2)$.
(iv): If $v_1, v_2 \in \{v \in V : v \notin V_t\}$ and $v_2 > v_1$, then $\min E^+(v_2) > \max E^+(v_1)$.

Every finite acyclic rooted directed graph is isomorphic to a canonical graph, and graph entropy is an isomorphism invariant; therefore, we concentrate on canonical graphs from now on.

Let $G = (V, E)$ be a finite rooted acyclic directed canonical graph. There is a unique mapping $\psi_G : V \to V_t^*$ satisfying the following two rules:

(i): $\psi_G(v) = v,\ v \in V_t$.

(ii): If $v \in V,\ v \notin V_t$, $r = \min E^+(v)$, and $s = \max E^+(v)$, then

$$\psi_G(v) = \psi_G(v(r)) * \psi_G(v(r+1)) * \cdots * \psi_G(v(s)),$$

where $v(e)$ denotes the vertex at which edge e terminates.

Let $(v_1, v_2, \ldots, v_k)$ be the string $\psi_G(1)$. If $(x_1, x_2, \ldots, x_k)$ is a string of the same length, whose symbols $\{x_i\}$ are selected from any set whatsoever, we write $G \to x$ if there is a one-to-one mapping $f : V_t \to \{x_1, x_2, \ldots, x_k\}$ such that $x_i = f(v_i)$ for $1 \le i \le k$. Let $\mathcal{G}$ denote the set of all finite rooted acyclic directed canonical graphs G such that ψ_G is one-to-one.

We fix a finite nonempty set A for the rest of the paper.

Theorem 1. *For each $x \in A^*$, let $G(x) = (V(x), E(x))$ be a graph in $\mathcal{G}$ such that $G(x) \to x$. Let $(X_1, X_2, \ldots)$ be an A-valued stationary ergodic process with entropy H. Assume that*

$$|E(X_1, X_2, \ldots, X_n)|/n \to 0 \text{ almost surely as } n \to \infty.$$

Then

$$H(G(X_1, X_2, \ldots, X_n))/n \to H \text{ almost surely as } n \to \infty.$$

3. Applications

1. For each $x \in A^*$, let $G(x) = (V(x), E(x))$ be a graph in $\mathcal{G}$ such that $G(x) \to x$. Suppose that $\max\{|E(x)| : x \in A^n\} = o(n)$. As shown in [2], there is a computationally attractive data compression algorithm that assigns to each sufficiently long $x \in A^*$ a binary codeword of length approximately equal to $H(G(x))$. Theorem 1 tells us that this algorithm optimally compresses the first n data samples generated by any stationary ergodic A-valued process, asymptotically as $n \to \infty$. ("Optimally compresses" refers to the well-known fact [1] that no compression algorithm can achieve an asymptotic compression rate in code bits per data sample less than the entropy of the process generating the data samples.)

2. The well-known Lempel-Ziv parsing rule [5] partitions each string $x \in A^*$ into $t = t(x)$ phrases such that

(i): Each phrase is either a singleton or is obtained by adjoining a symbol to the end of a preceding phrase.

(ii): The first $t - 1$ phrases are distinct.

For example, the Lempel-Ziv parsing of the data string 0010110111 is (0), (01), (011), (0111). Let $(X_1, X_2, \ldots)$ be an A-valued stationary ergodic process with entropy H. Theorem 1 can be used to deduce the asymptotic expansion

$$t(X_1, X_2, \ldots, X_n) = \frac{Hn}{\log n} + o\left(\frac{n}{\log n}\right) \text{ almost surely.}$$

(One defines a graph $G(x) \to x$ such that $H(G(x))$ is approximately equal to $t(x) \log n$ whenever $x \in A^n$ and n is large.)

3. Let $\phi : [0,\infty) \to (-\infty,\infty)$ be the function such that $\phi(0) = 0$ and $\phi(x) = x \log x$ for $x > 0$. Let $n > 1$ be an integer and let $x \in A^{2^n}$. For each integer k such that $0 \leq k \leq n$, let $S_k(x)$ be the set of all $y \in A^{2^k}$ that appear in the partitioning of x into substrings of length 2^k. If $y \in A^{2^k}$ for $0 \leq k < n$, let $N_l(y|x)$ be the number of z such that $y * z \in S_{k+1}(x)$ and let $N_r(y|x)$ be the number of z such that $z * y \in S_{k+1}(x)$. Define $Q_k(x)$ to be the number

$$Q_k(x) = \phi\left(\sum_{y \in A^{2^k}} \{N_l(y|x) + N_r(y|x) - 1\}\right) - \sum_{y \in A^{2^k}} \phi\left(N_l(y|x) + N_r(y|x) - 1\right).$$

Let $(X_1, X_2, \ldots)$ be an A-valued stationary ergodic process with entropy H. Using Theorem 1 one can deduce the following limit formula for H:

$$\frac{1}{2^n} \sum_{k=0}^{n-1} Q_k(X_1, X_2, \ldots, X_{2^n}) \to H \text{ almost surely as } n \to \infty.$$

Each string y lying in the union of the $S_k(x), 0 \leq k \leq n$, generates a vertex $v(y)$ of a graph $G(x)$. If the length of y is at least two, two edges emanate from $v(y)$, one going to $v(y_1)$ and one going to $v(y_2)$, where y_1 and y_2 are the right and left halves of y. One then applies Theorem 1 to the graphs $\{G(x)\}$.

4. Ancillary results

Lemma 1. *Let k be a positive integer. Let $\mathcal{G}_k = \{G = (V,E) \in \mathcal{G} : |E| \leq k\}$. The members of $\mathcal{G}_k$ can be assigned distinct binary codewords so that*

- **(i):** *The codeword assigned to $G \in \mathcal{G}_k$ is of length no greater than $H(G) + 6k + 1$.*
- **(ii):** *No codeword is a prefix of any other codeword.*

Lemma 2. *For $0 < \epsilon < 1$, and n sufficiently large, define $h_\epsilon(x)$ for $x \in A^n$ by*

$$h_\epsilon(x) = \min\{H(G) : G = (V,E) \in \mathcal{G},\ G \to x,\ |E| < n\epsilon\}.$$

Then

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} \log\left(\sum_{x \in A^n} 2^{-h_\epsilon(x)}\right) = 0.$$

Lemma 3. *Let k be a positive integer. Then there exists a function $f_k : (0,\infty) \to (-\infty,\infty)$ satisfying $\lim_{t \to 0+} f_k(t) = 0$ for which*

$$H(G)/n \leq f_k\left(|E|/n\right) - \frac{1}{nk} \sum_{i=0}^{n-k+1} \log \mu(x_i, x_{i+1}, \ldots, x_{i+k-1}) \tag{4.1}$$

whenever $n \geq k$, μ is a probability distribution on A^k, $x \in A^n$, and $G = (V,E) \in \mathcal{G}$ with $G \to x$.

5. Proofs

Proof of Lemma 1. We call the binary symbols in the codeword for $G = (V,E) \in \mathcal{G}_k$ "codebits". The first $6k$ codebits in the codeword serve to identify each of the following six entities concerning G:

- **(i):** $|E|$.
- **(ii):** $|V|$.
- **(iii):** V_t.

(iv): The cardinalities of the sets $E^+(v),\ v \in V,\ v \notin V_t$.
(v): The cardinalities of the sets $E^-(v),\ v \in V,\ v \neq 1$.
(vi): The positions in the vector $(v(1), v(2), \ldots, v(|E|))$ where each vertex $v \neq 1 \in V$ first appears.

Let s_G be the string obtained from $(v(1), v(2), \ldots, v(|E|))$ by deleting the first appearance of each vertex $v \neq 1 \in V$. The J remaining codebits identify the string s_G, and one then identifies G from s_G and items (i)–(vi). The string s_G lies in the set S_G of all strings of length $|E| - |V| + 1$ in which each $v \neq 1 \in V$ appears $|E^-(v)| - 1$ times. Taking $J = \lceil \log |S_G| \rceil$, $J \leq H(G) + 1$.

Proof of Lemma 2. Fix n so large that $h_\epsilon(x)$ is defined for each $x \in A^n$. For each such x, pick a $G_x = (V_x, E_x) \in \mathcal{G}$ such that $H(G_x) = h_\epsilon(x)$, $G_x \to x$, and $|E_x| < n\epsilon$. According to Lemma 1, we can assign to each $G \in \mathcal{G}_n = \{G_x : x \in A^n\}$ a binary codeword of length $L(G) \leq H(G) + 6n\epsilon + 1$. Kraft's inequality from information theory ([1], page 82) tells us that $\sum_{G \in \mathcal{G}_n} 2^{-L(G)} \leq 1$. From this and the fact that there are $\leq |A|^{n\epsilon}$ strings $x \in A^n$ such that $G \to x$ for each $G \in \mathcal{G}_n$, it follows that

$$\sum_{x \in A^n} 2^{-L(G_x)} \leq |A|^{n\epsilon},$$

and therefore

$$\sum_{x \in A^n} 2^{-h_\epsilon(x)} \leq 2^{6n\epsilon+1} |A|^{n\epsilon}.$$

Proof of Lemma 3. Fix k and a probability distribution μ on A^k. For each pair j, n in which $0 \leq j < k$ and $n \geq k$, let $W_{j,n} = \{1 \leq i \leq n - k + 1 : i \equiv j \bmod k\}$. For each string $s = (s_1, s_2, \ldots, s_n) \in A^*$, define $\lambda(s)$ as follows:

(i): $\lambda(s) = 1,\ n < k$.
(ii): $\lambda(s) = \max_{0 \leq j < k} \left[\prod_{i \in W_{j,n}} \mu(s_i, s_{i+1}, \ldots, s_{i+k-1}) \right],\ n \geq k,$

where an empty product is taken to be one. For each $s \in A^*$, define $\lambda^*(s) = C^{-1} |s|^{-2} \lambda(s)$, where $|s|$ denotes the length of s and C is the positive real constant (depending on k) that makes the numbers $\{\lambda^*(s) : s \in A^*\}$ sum to one. Fix $n \geq k$, $x = (x_1, \ldots, x_n) \in A^n$, and $G = (V, E) \in \mathcal{G}$ such that $G \to x$. Let $r = |E| - |V| + |V_t| + 1$. The function which carries $\psi_G(1)$ into x also carries each $\psi_G(v)$ into a string $\psi_G^*(v)$ $(v \in V,\ v \neq 1)$. There exist strings $\{s_i : 1 \leq i \leq r\}$ such that

(i): $\{s_i : |E| - |V| + 1 < i \leq r\} = \{x_1, x_2, \ldots, x_n\}$.
(ii): $|\{1 \leq i \leq |E| - |V| + 1 : s_i = \psi_G^*(v)\}| = |E^-(v)| - 1,\ v \in V, v \neq 1$.
(iii): $x = \tilde{s}_1 * \tilde{s}_2 * \cdots * \tilde{s}_r$, for some permutation $\{\tilde{s}_i\}$ of $\{s_i\}$.

Note that

$$\prod_{i=1}^{n-k+1} \mu(x_i, \ldots, x_{i+k-1}) \leq \left[\prod_{i=1}^{r} \lambda(s_i) \right]^k.$$

Replacing $\lambda(s_i)$ by $C|s_i|^2 \lambda^*(s_i)$, and taking the logarithm of both sides,

$$-\sum_{i=1}^{|E|-|V|+1} \log \lambda^*(s_i) \leq 2 \sum_{i=1}^{r} \log |s_i| + r \log C - \frac{1}{k} \sum_{i=1}^{n-k+1} \log \mu(x_i, \ldots, x_{i+k-1}).$$

Concavity of the logarithm function yields the bound

$$2\sum_{i=1}^{r}\log|s_i| \le 2r\log\left(\frac{n}{r}\right).$$

For each $s = \psi_G^*(v)$, $v \in V$, $v \neq 1$, define $\sigma(s) = (|E^-(v)|-1)/(|E|-|V|+1)$. Then

$$H(G) = -\sum_{i=1}^{|E|-|V|+1}\log\sigma(s_i) \le -\sum_{i=1}^{|E|-|V|+1}\log\lambda^*(s_i).$$

Condition (4.1) is true with $f_k(t) = \sup_{0<\delta\le t}\{\delta\log C - 2\delta\log\delta\}$.

Proof of Theorem 1. Let $(X_1, X_2, \dots)$ be a stationary ergodic A-valued process with entropy H. For $n \ge 1$, let $X^n = (X_1, X_2, \dots, X_n)$. If $x \in A^*$ has length n, define $\mu(x) = \Pr[X^n = x]$. From (4.1) we see that $\limsup_{n\to\infty} H(G(X^n))/n \le H$ almost surely.

Fix $\delta > 0$. By Lemma 2, there exists ϵ such that for n sufficiently large,

$$\Pr\left[\log\left(\frac{2^{-h_\epsilon(X^n)}}{\mu(X^n)}\right) \ge n\delta\right] < 2^{-n\delta/2}.$$

Applying the Borel-Cantelli lemma,

$$\limsup_{n\to\infty}\frac{1}{n}\log\left(\frac{2^{-h_\epsilon(X^n)}}{\mu(X^n)}\right) \le \delta \text{ almost surely.} \tag{5.1}$$

Replacing $h_\epsilon(X^n)$ by $H(G(X^n))$ in (5.1), and then using the fact that

$$-n^{-1}\log\mu(X^n) \to H \text{ a.s.}$$

(Shannon-McMillan-Breiman Theorem), one concludes that

$$\liminf_{n\to\infty} H(G(X^n))/n \ge H - \delta \text{ almost surely,}$$

for an arbitrary $\delta > 0$.

References

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley, New York, 1991. MR **92g:**94001

[2] J. Kieffer and E.-H. Yang, "A complexity reduction method for lossless source code design," Technical Report, Dept. of Electrical Engineering, University of Minnesota Twin Cities, 1995 (http://www.ee.umn.edu/users/kieffer).

[3] J. Kieffer, E.-H. Yang, G. Nelson, and P. Cosman, "Lossless compression via bisection trees," Technical Report, Dept. of Electrical Engineering, University of Minnesota Twin Cities, 1996 (http://www.ee.umn.edu/users/kieffer).

[4] J. Kieffer, G. Nelson, and E.-H. Yang, "Tutorial on the quadrisection method for lossless data compression," Technical Report, Dept. of Electrical Engineering, University of Minnesota Twin Cities, 1996 (http://www.ee.umn.edu/users/kieffer).

[5] A. Lempel and J. Ziv, *On the complexity of finite sequences*, IEEE Trans. Inform. Theory, **22** (1976), 75–81. MR **52:**10234

Department of Electrical Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455

E-mail address: `kieffer@ee.umn.edu`

Department of Mathematics, Nankai University, Tianjin 300071, P. R. China

E-mail address: `ehyang@irving.usc.edu`