THE ENTROPY OF FUNCTIONS OF FINITE-STATE
MARKOV CHAINS:

DAVID BLACKWELL
UNIVERSITY OF CALIFORNIA, BERKELEY, CALIFORNIA

0. Summary. It is shown that the entropy $H$ of an ergodic stationary process $\{y_n, -\infty < n < \infty\}$ with states $a = 1, \ldots, A$ such that $y_n = \phi(x_n)$, where $\{x_n\}$ is a stationary ergodic finite-state Markov process with states $i = 1, \ldots, I$ and transition matrix $M = [m(i, j)]$ is given by

$$H = - \sum_{a} r_a(w) \log r_a(w) dQ(w),$$

where $r_a$ is a function, defined on the set $W$ of all $w = (w_1, \ldots, w_i)$ such that $w_i \geq 0$, $w_i = 1$, by $r_a(w) = \sum_{i, j(i) = 3} w_i m(i, j)$ and $Q$ is the distribution of the conditional distribution of $x_0$ given $y_0, y_{-1}, \ldots$. An integral equation is obtained for $Q$, and a method is given for showing, under rather strong hypotheses, that the solution of this integral equation is unique. An example in which $Q$ is singular is given.

1. Introduction. Let $\{x_n, -\infty < n < \infty\}$ be a stationary ergodic stochastic process with a finite set of states $i = 1, \ldots, I$. For any finite sequence $s = (i_1, \ldots, i_k), 1 \leq i_1, \leq I, k = 1, 2, \ldots$, let $q(s) = P(x_1 = i_1, \ldots, x_k = i_k)$, and let $z_k = q(x_1, \ldots, x_k)$. In developing Shannon's fundamental work [3], McMillan [2] has shown that the sequence $\{z_n\}$ of random variables is asymptotically constant in the sense that there is a constant $H \geq 0$, called the entropy of the process $\{x_n\}$, such that $E[n^{-1} \log z_n + H] \to 0$ as $n \to \infty$. This result, which implies that for large $n$ it is very probable that the sequence of values of $x_1, \ldots, x_n$ which actually occurs is one whose probability is about $2^{-Hn}$, where the log above has base 2, is fundamental for the development of information theory. Moreover, if $u(i) = P(x_1 = i | x_0, x_{-1}, \ldots)$ and $v = u(x_1)$, McMillan shows that

$$H = - E \log v.$$
If \( \{x_n\} \) is a Markov process with
\[
\lambda_i = P(x_n = i), \quad m(i, j) = P(x_{n+1} = j \mid x_n = i, x_{n-1}, \ldots),
\]
we have \( u(i) = m(x_0, i) \), \( v = m(x_0, x_1) \), so that
\[
H = - \sum_{ij} \lambda_i m(i, j) \log m(i, j).
\]
Thus the entropy of a Markov process is easily calculated. However, if \( \Phi \) is
a function defined on 1, \ldots, \( I \) with values 1, \ldots, \( A \), no formula for the entropy
of \( \{y_n = \Phi(x_n)\} \), comparable in simplicity to (3), is known. This entropy is
clearly a function of the matrix \( M = \|m(i, j)\| \) and the function \( \Phi \). In this
paper we study the entropy of the \( \{y_n\} \) process; our result suggests that this
entropy is intrinsically a complicated function of \( M \) and \( \Phi \).

2. The main result. Our main result is the following.

**Theorem 1.** Let \( \{x_n \mid -\infty < n < \infty\} \) be a stationary ergodic Markov process
with states \( i = 1, \ldots, I \) and transition matrix \( M = \|m(i, j)\| \). Let \( \Phi \) be a function
defined on 1, \ldots, \( I \) with values \( \alpha = 1, \ldots, A \) and let \( y_n = \Phi(x_n) \). The entropy of
the \( \{y_n\} \) process is given by,
\[
H = - \int \sum_{a} r_a(w) \log r_a(w) dQ(w),
\]
where \( Q \) is a probability distribution on the Borel sets of the set \( W \) of vectors \( w = (w_1, \ldots, w_i) \) with \( w_i \geq 0, \sum_i w_i = 1 \) and \( r_a(w) = \sum_i \sum_{j \neq \Phi(i)} w_i m(i, j) \). The
distribution \( Q \) is concentrated on the sets \( W_1, \ldots, W_A \), where \( W_a \) consists of all
\( w \in W \) with \( w_i = 0 \) for \( \Phi(i) \neq a \) and satisfies the equation
\[
Q(E) = \sum_a \int_{f_a^{-1}(E)} r_a(w) dQ(w),
\]
where \( f_a \) maps \( W \) into \( W_a \), with the \( j \)th coordinate of \( f_a(w) \) given by \( \sum_j w_j m(i, j) r_a(w) \)
for \( \Phi(j) = a \).

**Proof.** Let \( x_n = (x_{n-1}, \ldots, x_1) \) be the conditional distribution of \( x_n \) given
\( y_n, y_{n-1}, \ldots, y_1 \), that is, \( x_n = P(x_n = i \mid y_n, y_{n-1}, \ldots) \). If \( u(a) = P(y_1 = a \mid y_0, y_{-1}, \ldots) \)
and \( v = u(y_1) \), we shall show that
\[
E(\log v \mid x_n = w) = \sum_a r_a(w) \log r_a(w),
\]
so that, if \( Q \) is the distribution of \( x_0 \), (1) holds. We shall also show that \( \{x_n\} \) is
a stationary Markov process, with
\[
P(x_{n+1} \in E \mid x_n, x_{n-1}, \ldots) = \sum_{a \in E} r_a(x_n),
\]
so that (4) is simply the equation satisfied by the stationary distribution \( Q \) of
the \( \{x_n\} \) process.
The sequence \( \{x_n\} \) is easily seen to be stationary, for say \( x_{ni} = \delta_s(y_0, y_{-1}, \ldots) \), let \( \beta \) be any bounded random variable depending only on \( y_n, y_{n-1}, \ldots \), say \( \beta = \psi(y_n, y_{n-1}, \ldots) \), and let \( \gamma(i) = 1 \), \( \gamma(j) = 0 \) for \( j \neq i \). Then
\[
E \beta \gamma(x_n) = E \psi(y_0, y_{-1}, \ldots) \gamma(x_n) = E \psi(y_0, y_{-1}, \ldots) g_i(y_0, y_{-1}, \ldots) =
E \beta g_i(y_n, y_{n-1}, \ldots),
\]
so that \( \alpha_{ni} = g_i(y_n, y_{n-1}, \ldots) \). Thus \( \alpha_n = g(y_n, y_{n-1}, \ldots) \), with \( g = (g_1, \ldots, g_t) \), and \( \{x_n\} \) is stationary.

To establish (6) we have
\[
P(x_{n+1} = i \mid y_n, y_{n-1}, \ldots) = \sum_j P(x_n = j, x_{n-1} = i \mid y_n, y_{n-1}, \ldots) = \sum_j \alpha_{nj} m(j, i).
\]
Say \( \Phi(i) = a \). Summing (8) over all \( i \) for which \( \Phi(i) = a \) yields
\[
P(y_{n+1} = a \mid y_n, y_{n-1}, \ldots) = r_a(\alpha_n).
\]
But also, using (9)
\[
P(x_{n+1} = i \mid y_n, y_{n-1}, \ldots) = P(y_{n+1} = a \mid y_n, y_{n-1}, \ldots) P(x_{n+1} = i \mid y_{n+1} = a, y_n, y_{n-1}, \ldots) = \beta_a(\alpha_n) \beta_{ni},
\]
where \( \beta_a = 1 \) if \( y_{n+1} = a \), \( 0 \) if \( y_{n+1} \neq a \). Equations (8) and (10) yield, for \( y_{n+1} = a = \Phi(i) \),
\[
\alpha_{n+1, i} = \sum_j \alpha_{nj} m(j, i) / r_a(\alpha_n),
\]
so that
\[
\alpha_{n+1} = f_{\Phi, x_n}(\alpha_n).
\]
Since \( f_{\Phi}(w) \in \mathcal{W}_a \) for all \( w \), the distribution \( Q \) of \( x_n \) is concentrated on \( \mathcal{W}_1, \ldots, \mathcal{W}_A \).
From (12) we obtain
\[
P(x_{n+1} \in E \mid x_n, x_{n-1}, \ldots) = \sum_{a \in f_{\Phi}(\alpha_n) \cap E} P(y_{n+1} = a \mid x_n, x_{n-1}, \ldots),
\]
so that, using (9), we obtain (6). Taking expectations in (6) yields (4).

Finally, to compute the entropy of \( \{y_n\} \), we have \( \omega(a) = P(y_1 = a \mid y_0, y_{-1}, \ldots) = r_a(\alpha_0), \nu = \omega(y_1) = r_0(\alpha_0) \), so that
\[
E(\log \omega) = \sum_a \omega(a) \log r_a(\alpha_0) = \sum_a r_a(\alpha_0) \log r_a(\alpha_0).
\]
Thus (5) is established, and the proof is complete.

3. Singularity of \( Q \). The distribution \( Q \) may be concentrated on a finite set; for instance if \( \Phi \) is the identity function, \( Q \) is concentrated on the \( I \) unit vectors. Also \( Q \) may be concentrated on a countable set. Whenever there is an \( a \) for
which \( \varphi(i) = a \) for exactly one \( i \), \( Q \) will be concentrated on a set which is at most countable, for \( x_n = \delta_i \), the \( i \)th unit vector, whenever \( y_n = a \), so that all values of \( x_k \) have the form \( f_{a_k}(f_{a_k-1}(\ldots(f_{a_1}(\delta_i))\ldots)) \) for some finite sequence \( a_1, \ldots, a_k \). Thus for

\[
M = \begin{pmatrix}
0 & 1 & 2 \\
3 & 4 & 5 \\
6 & 0 & 0 \\
\end{pmatrix}
\]

with \( \Phi(1) = 1, \Phi(2) = \Phi(3) = 2 \),

if \( y_n = 1 \),

\[
\alpha_n(y_n, y_{n-1}, \ldots) = (1, 0, 0) = \delta_i
\]

if \( y_n = \ldots = y_{n-k+1} = 2, y_{n-k} = 1 \).

We have \( P(x_0 = \delta_i) = P(x_0 = 1) = 2/7 \)

\[
P(x_0 = w_k) = P(x_0 = 1, x_1 = \ldots = x_k = 2 \text{ or } 3) = 2(2^k + 1)/7.
\]

Also \( r_i(\delta_i) = 0, r_2(\delta_i) = 1 \),

\[
r_i(w_k) = (2^k + 2)/3(2^k + 1) = b_k, \\
\]

so that

\[
\mathcal{H} = - \sum_{k=1}^{m} p_k(b_k \log b_k + (1 - b_k) \log (1 - b_k)).
\]

Also, \( Q \) may be continuous. We conjecture that, when \( Q \) is continuous, it is generally singular in the following sense. Let \( q_0(\vec{y}) = Q(\vec{y} \cap \mathcal{W}_a) \), so that \( Q_0 \) is concentrated on the set \( \mathcal{W}_a \) of dimension \( d_a \), say, where \( d_a + 1 \) is the number of \( i \) for which \( \varphi(i) = a \). If each \( d_a \geq 1 \) and \( Q_0 \) is continuous and concentrated on a subset \( S_a \) of \( \mathcal{W}_a \) whose \( d_a \)-dimensional Lebesgue measure is zero, we shall say that \( Q \) is singular.

It can be shown that, whenever each \( d_a \geq 1 \), all rows of \( M \) are nearly identical, that is, \( \max m(i, j) - \min m(i, j) \) is sufficiently small, and no element of \( m(i, j) \) is too near zero, each \( Q_a \) is concentrated on a set of \( d_a \)-dimensional Lebesgue measure zero. We illustrate the method with the example:

\[
M = \begin{pmatrix}
\epsilon + c & c & -\epsilon & c \\
c & c - \epsilon & c & c + \epsilon \\
-\epsilon & c & c + \epsilon & c \\
c & c + \epsilon & c - \epsilon & c \\
\end{pmatrix}
\]

where \( c = \frac{1}{2}, 0 < \epsilon < c \), with

\[
\Phi(1) = \Phi(2) = 1, \quad \Phi(3) = \Phi(4) = 2.
\]
Write \( w_1(x) = (x, 1 - x, 0, 0), w_2(x) = (0, 0, x, 1 - x) \), so that \( W \) consists of all \( w_a(x), 0 \leq x \leq 1 \). Write, for any Borel set \( B \) of \((0, 1),\)

\[ m_a(B) = Q_a(w_aB). \]

We shall show that \( m_1, m_2 \) are continuous and concentrated on sets of linear Lebesgue measure 0, if \( \varepsilon \) is sufficiently small. We have

\[
\begin{align*}
r_1(w_1(x)) &= 2c - \varepsilon + 2x = r_2(w_2(x)) = R(x), \text{ say}, \\
r_1(w_2(x)) &= 2c + \varepsilon - 2x = r_2(w_1(x)) = r(x), \text{ say}.
\end{align*}
\]

\[
\begin{align*}
f_1(w_1(x)) &= w_1((c + \varepsilon)/R(x)), \\
f_1(w_2(x)) &= w_2((c - \varepsilon)/r(x)), \\
f_2(w_1(x)) &= w_2((c - \varepsilon)/r(x)).
\end{align*}
\]

Write \( F(x) = (c + \varepsilon)/R(x), f(x) = (c - \varepsilon)/r(x) \). The equation (4), when expressed in terms of \( m_1, m_2 \), becomes

\[
m_1(B) = \int_{F^{-1}B} R(x) \, dm_1 + \int_{f^{-1}B} r(x) \, dm_2,
\]

\[
m_2(B) = \int_{F^{-1}B} R(x) \, dm_2 + \int_{f^{-1}B} r(x) \, dm_1.
\]

Thus if \( m_1, m_2 \) are concentrated on \( S_1, S_2 \) they are also concentrated on \( FS_1 \cup fS_2, fS_1 \cup FS_2 \), respectively, But if \( \mu \) denotes Lebesgue measure,

\[
\mu(FS_1 \cup fS_2) \leq \max_{0 \leq \varepsilon \leq 1} |F'(x)| \mu(S_1) + \max_{0 \leq \varepsilon \leq 1} |f'(x)| \mu(S_2),
\]

and a similar inequality holds for \( \mu(fS_1 \cup FS_2) \). We have

\[
\max_{0 \leq \varepsilon \leq 1} |F'(x)| = \max_{0 \leq \varepsilon \leq 1} |f'(x)| = -|F'(0)| = -f'(1) = \varepsilon^2/(2c - \varepsilon)^2,
\]

so that

\[
\max_{0 \leq \varepsilon \leq 1} (\mu(FS_1 \cup fS_2), \mu(fS_1 \cup FS_2)) \leq \max (\mu(S_1), \mu(S_2)) \frac{2\varepsilon^2}{(2c - \varepsilon)^2}
\]

Thus if \( 2\varepsilon^2/(2c - \varepsilon)^2 < 1 \), that is, \( \varepsilon < (\sqrt{2} - 1)/2 = .207... \), \( m_1, m_2 \) are concentrated on sets of arbitrarily small Lebesgue measure and consequently on sets of Lebesgue measure zero.

Finally, to see that \( m_1, m_2 \) are continuous, let \( \lambda_a(x) = m_a((a, x)) \), \( \lambda_a(x) = \max \lambda_a(x). \) Since \( F \) and \( f \) are strictly monotone decreasing and \( F(x) > 2c > f(x) \) for all \( x \), at most one of \( f^{-1}w_a, F^{-1}w_a \) is nonempty, and this consists of a single point. (15) yields

\[
\max (\lambda_1(x_1), \lambda_2(x_2)) \leq (2c + \varepsilon) \max (\lambda_1(x_1), \lambda_2(x_2))
\]

so that \( \lambda_a(x) \equiv 0, \lambda_a(x) \equiv 0, \) and \( m_1, m_2 \) are continuous.

The above argument showing that the \( Q_\varepsilon \) are concentrated on sets of Lebesgue measure zero depended only on the fact that the derivatives of the functions \( f_a \) are small. These derivatives will be small whenever the rows of \( M \) are nearly identical (if the rows of \( M \) are identical, the \( f_a \) are constant) and no element is very near 0, and the argument can be carried through.
4. Uniqueness of solution of (4). Even if we find a solution $Q$ of (4), we cannot conclude that this $Q$ is the distribution of $\alpha_n$ unless the solution of (4) is unique.

Again, if the rows of $M$ are nearly identical and no element is very near zero, (4) can be shown to have a unique solution, using a method exploited by Harris [1]. We sketch the proof of the following theorem.

**Theorem 2.** Let $f_1, \ldots, f_d$ be functions defined on a bounded metric space $W$ with metric $\sigma$ such that, for some $q < 1$

$$\sigma((f_d(w_1), f_d(w_2)) \leq q \sigma(w_1, w_2)$$

for all $w_1, w_2, w_3$, and let $r_1(w), \ldots, r_d(w)$ satisfy

$$\sum_{a} r_a(w) = 1, \quad r_a(w) \geq \varepsilon > 0$$

for all $w$, and

$$|r_a(w_1) - r_a(w_2)| \leq c_0 \sigma(w_1, w_2)$$

for some $c_0 > 0$.

Then there is at most one probability distribution $Q$ on the Borel sets of $W$ such that for every Borel set $E$

$$Q(E) = \sum_{f_n \in E} \int r_a(w) dQ(w).$$

**Proof.** Let $h_1, h_2, \ldots$ be independent variables, each uniformly distributed on $(0, 1)$, let $\alpha_0, \beta_0$ be independent of each other and of $\{h_n\}$ with arbitrary distributions $Q_0, Q_1$ on $W$, and define, inductively, $a_{n+1}, b_{n+1}$ as the integers $a, b$ for which

$$r_1(\alpha_n) + \ldots + r_{a-1}(\alpha_n) \leq h_{n+1} < r_1(\alpha_n) - \ldots - r_d(\alpha_n),$$

$$r_1(\beta_n) + \ldots + r_{b-1}(\beta_n) \leq h_{n+1} < r_1(\beta_n) - \ldots - r_d(\beta_n),$$

$$\alpha_{n+1} = f_{a_{n+1}}(\alpha_n), \quad \beta_{n+1} = f_{b_{n+1}}(\beta_n).$$

Thus $\{\alpha_n\}, \{\beta_n\}$ are Markov processes with the same transition law $w \rightarrow f_a(w)$ with probability $r_a(w)$, $a = 1, \ldots, d$ and initial distributions $Q_0, Q_1$ respectively. We show that, with probability 1, $a_n = b_n$ for all sufficiently large $n$. Fix $n$, and let

$$\Pi(k) = P(a_{n+i} = b_{n+i}, 1 \leq i \leq k \mid \alpha_0, \beta_0, h_1, \ldots, h_n).$$

Then

$$\Pi(k + 1) = \Pi(k) P(a_{n+k+1} = b_{n+k+1} \mid a_{n+i} = b_{n+i}, 1 \leq i \leq k, \alpha_0, \beta_0, h_1, \ldots, h_n).$$

Now $a_{n+i} = b_{n+i}, 1 \leq i \leq k$ implies from (16) $\sigma(\alpha_{n+k}, \beta_{n+k}) \leq q^d$, so that from (17),

$$|r_a(\alpha_{n+k}) - r_a(\beta_{n+k})| \leq q^d,$$

where $d = \sup_{w_1, w_2} \sigma(w_1, w_2)$, and

$$|r_1(\alpha_{n+k}) + \ldots + r_d(\alpha_{n+k}) - r_1(\beta_{n+k}) - \ldots - r_d(\beta_{n+k})| \leq Acd_2^k.$$
\[ i \leq n + k \text{, falls in one of the } A \text{ intervals } (\sum r(x_{i+k}), \sum r(\beta_{n+k})), \text{ each of whose lengths, when } x_{i+k} = \beta_{n+k}, 1 \leq i \leq k, \text{ is at most } Acdg^k. \text{ Thus the last factor in (19) is at least } 1 - A^2cdg^k, \text{ and}
\]
\[ \prod (k + 1) \geq \prod (1 - A^2cdg^k). \]

Also, for every \( k \),
\[(21) \quad \prod (k) \geq P(x_{i+k} = b_{i+k} = 1, 1 \leq i \leq k | x_0, \beta_0, h_1, \ldots, h_n) \geq e^k \]
from (17). Choose \( k_0 \) so large that \( Acdg^{k_0} < 1 \). Then, from (20) and (21),
\[ \prod (k) \geq \frac{e^k}{k_0} \prod (1 - Acdg^k) = \delta > 0 \]
for all \( k \). Thus the event \( a_n = b_n \) for sufficiently large \( n \), which depends only on \( x_n, \beta_n, h_1, \ldots, h_n \), has conditional probability at least \( \delta > 0 \) for all \( x, \beta, h_1, \ldots, h_n \), so that it has probability one.

From (16), \( a_n = b_n \) for sufficiently large \( n \) implies \( o(x_n, \beta_n) \to 0 \) as \( n \to \infty \), so that
\[(22) \quad P(o(x_n, \beta_n) \to 0 \text{ as } n \to \infty) = 1. \]

Suppose each of \( Q_0, Q_1 \) satisfies (4). Then \( x_n \) has distribution \( Q_0 \) and \( \beta_n \) has distribution \( Q_1 \) for all \( n \) so that, from (22), \( Q_0 = Q_1 \).

We may apply the theorem to the example considered above, as follows. Let \( W = W_1 \cup W_2 \) and define \( o(w_1, w_2) = |w_1 - w_2| \) if \( w_1, w_2 \in W_1 \) or \( w_1, w_2 \in W_2 \), \( o(w_1, w_2) = 10 \) if \( w_1 \in W_1, w_2 \in W_2 \). For \( v_1, v_2 \in W_1 \), say \( v_1 = v_1(x), v_2 = v_1(y) \), we have
\[ o(f_1(v_1), f_1(v_2)) = \sqrt{2} |F(x) - F(y)| \leq \sqrt{2} \max_{t \in (0,1)} |F(t)|x - y = \]
\[ \left( \frac{\xi c}{2c - \xi} \right)^2 o(v_1, v_2), \]
so that (16) is satisfied for \( v_1, v_2 \in W_1, \alpha = 1 \) for \( \left( \frac{\xi c}{2c - \xi} \right)^2 \leq 1, \) that is \( \xi < \frac{1}{4} \).

Similarly for \( \alpha = 2, \) or \( v_1, v_2 \in W_2, \) or both. For \( v_1 \in W_1, v_2 \in W_2 \)
\[ o(f_0(v_1), f_0(v_2)) \leq \max_{v \in W_1} o(v, v') = \sqrt{2}, \]

while \( o(v_1, v_2) = 10, \) so that (16) is satisfied for all cases. Also \( \min_x r_1(w) = \min_x r_2(w) = \min_x R(x) - \min_x r(x) = 2c - \xi, \) so that \( r_0(w) \) is bounded away from 0 for all \( \xi < \frac{1}{4}, \) so that (17) is satisfied. Finally (18) is clearly satisfied for any \( \xi. \) Thus for \( 0 < \xi < c, \) the system (15) has a unique solution \( m_1, m_2 \) such that \( m_1(E) + m_2(E) = 1 \) when \( E \) is the unit interval.

As stated above, for any matrix \( M \) whose rows are nearly identical and no element of which is very small, Theorem 2 can be used to show that (4) has a unique solution. The writer believes that (4) has a unique solution whenever \( M \) is indecomposable.
5. Two examples. If \( \{x_n\} \) is any ergodic stationary process with states \( i = 1, \ldots, I \) and \( y_n = \Phi(x_n) \), then

\[
0 \leq H_y \leq H_x,
\]

where \( H_x, H_y \) are the entropies of \( \{y_n\}, \{x_n\} \) respectively. This is obvious, since the random variables \( z_{x_k} = P(x_1, \ldots, x_k), z_{y_k} = P(y_1, \ldots, y_k) \) satisfy \( z_{x_k} \leq z_{y_k} \)

\[
n^{-1} \log z_{x_k} \rightarrow -H_x, n^{-1} \log z_{y_k} \rightarrow -H_y \text{ in mean. It can happen that } H_y = H_x \text{ even though } x_n \text{ is not a function of } y_n \text{ or even of } \{y_n\}.
\]

An example is given by the Markov process \( \{x_n\} \) with transition matrix

\[
M = \begin{pmatrix}
0 & a & 0 \\
0 & 0 & a \\
a & 0 & 0 \\
a & 0 & 0
\end{pmatrix},
\]

where \( a = \frac{1}{2} \), with \( \Phi(1) = \Phi(2) = 1, \Phi(3) = \Phi(4) = 2 \). A direct computation of \( H_x \) from (2) shows \( H_x = 1 \). The \( y_n \) are independent with values 1, 2 with probability \( \frac{1}{2} \) each, so that \( H_y = 1 \) also.

If the matrix \( M \) of a stationary ergodic Markov process \( \{x_n\} \) is aperiodic, and \( y_n = \Phi(x_n) \), then \( H_y = 0 \) only if \( \Phi \) is constant. However, there do exist stationary processes \( \{y_n\} \) of entropy 0 even though \( y_n \) is not a function of any finite segment \( y_{-1}, \ldots, y_{-m} \) of the past. A familiar example is the sequence \( y_n = \Phi(x_n) \), where \( \Phi(t) = 0 \) for \( 0 \leq t < \frac{1}{2}, \Phi(t) = 1 \), for \( \frac{1}{2} \leq t < 1 \), and \( x_n = \Theta + n\alpha \text{ mod } 1 \), where \( \Theta \) is uniform on \( 0 \leq t \leq 1 \) and \( \alpha \) is a fixed irrational number. The random variable \( \Theta \) is a function of \( y_0, y_1, \ldots \), for if \( t_1 \neq t_2 \), say, \( 0 < t_2 - t_1 = \Delta < \frac{1}{2} \), there is a \( k \leq 0 \) for which \( \frac{1}{2} - \Delta < t_1 + \frac{k\alpha}{2} \), so that \( 0 < t_2 + k\alpha < \frac{1}{2} + \Delta \), and \( y_k \) for \( \Theta = t_1 \) differs from \( y_k \) for \( \Theta = t_2 \). If \( t_2 - t_1 = \frac{1}{2} \), \( y_k \) for \( \Theta = t_1 \) differs from \( y_k \) for \( \Theta = t_2 \) for all \( k \). Thus \( \Theta \) is a function of \( y_0, y_1, \ldots \). Consequently \( y_1 \) is a function of \( y_0, y_1, \ldots \), and the process has entropy zero.