Simple Proofs of Some Theorems on Noiseless Channels

IMRE CSISZÁR

Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary

Shannon's capacity formula for memoryless and finite-state noiseless channels is proved in a simple elementary way, for arbitrary symbol costs; actually, a somewhat stronger result is proved. Further, a simple proof of a version of the noiseless coding theorem is given, based on the properties of entropy and avoiding combinatorial arguments; also, in the familiar proof of the mentioned theorem, a possible simplification is pointed out. Finally, a nearly optimal encoding for finite-state noiseless channels is suggested.

In this paper, the term channel will always mean a noiseless channel. \( Y = \{y_j\}_{j \in J} \) will denote a finite or countably infinite alphabet, where \( J = \{1, \cdots, m\} \) \((m \geq 2)\), or \( J = \{1, 2, \cdots\}\). The set of all finite sequences (including the void sequence) of letters of \( Y \) will be denoted by \( Y^* \).

1. A memoryless channel with alphabet \( Y \) is a device capable of transmitting arbitrary sequences from \( Y^* \), the cost of transmission of each letter depending on the transmitted letter only. The cost of transmission of \( y_j \in Y \) will be denoted by \( t_j \). We assume

\[
t^* = \inf_{j \in J} t_j > 0.
\]

(1.1)

If \( t \geq 0 \), let \( E(t) \) be the set of sequences \( u \in Y^* \) of cumulative cost just not exceeding \( t \), i.e.

\[
u = y_{j_1} \cdots y_{j_n} \in E(t) \iff \sum_{k=1}^{n} t_{j_k} \leq t, \quad \sum_{k=1}^{n} t_{j_k} + t^* > t. \quad (1.2)
\]

Let \( N(t) \) denote the number of sequences belonging to \( E(t) \) if \( t \geq 0 \); for \( t < 0 \) set \( N(t) = 0 \).

† This research was done while the author was visiting professor at the University Erlangen-Nürnberg, Erlangen, West Germany.
Following Shannon and Weaver (1948), we call the limit

\[ C = \lim_{t \to \infty} \frac{\log_2 N(t)}{t} \]  

the \textit{capacity} of the memoryless channel.

Observe that if \( \bar{N}(t) \) denotes the number of all sequences \( u = y_1 \cdots y_n \in Y^* \) with cumulative cost not exceeding \( t \) i.e. with \( \sum_{k=1}^{n} t_{jk} \leq t \) then (1.3) implies \( \lim_{t \to \infty} \log_2 \bar{N}(t)/t = C \) as well.

Furthermore, let \( N_0(t) \) denote the number of different sequences \( u \in Y^* \) of cumulative cost exactly \( t \). Actually, in Shannon's original definition of capacity, the role of \( N(t) \) has been played by \( N_0(t) \). However, as \( N_0(t) = 0 \) except for a countable set of values of \( t \), with that interpretation of \( N(t) \) the \( \lim \) in (1.3) ought to be replaced by \( \lim \sup \), cf. Krause (1962). We shall write

\[ C_0 = \lim_{t \to \infty} \sup \frac{\log_2 N_0(t)}{t} \]  

(1.3')

Apparently, Shannon had integral symbol costs with greatest common divisor 1 in mind, with \( t \) running through the integers only. For that case, he has derived the capacity formula (1.7) below. His argument, based on solving the difference equation (1.8) below, applies equally well for both \( C_0 \) and \( C \), and, moreover, it proves even that \( N_0(t)w_0^{-t} \) (as well as \( N(t)w_0^{-t} \)) converges to a finite limit as \( t \to \infty \), provided that the alphabet \( Y \) is finite; a closer study of the difference equation (1.8) (See De Bruijn and Erdős, 1951) enables one to obtain a similar (and even a sharper) result for infinite alphabets, as well.

For the case of arbitrary symbol costs, (1.3) seems to be a more adequate definition of capacity than (1.3'). Nevertheless, a rigorous proof of Shannon's capacity formula (1.7) for arbitrary symbol costs, on the basis of the definition (1.3'), has been given by Krause (1962). He has used the theory of Dirichlet series. The proof we are going to give for Proposition 1.1 below has the merit of being completely elementary. As to the equivalence of both definitions of capacity, see the remark after the proof of Proposition 1.2.

Set

\[ a(w) = \sum_{i \in J} w^{-i}. \]  

(1.4)

In the case \( J = \{1, 2, \cdots \} \) let \( w^* \geq 1 \) denote the infimum of the values \( w \) for which the series (1.4) converges; if it diverges for every positive \( w \), let \( w^* = +\infty \). In the case \( J = \{1, \cdots, m\} \), we set \( w^* = 1 \).
The function \( a(w) \) is clearly continuous and monotonically decreasing in the interval \( (w^*, +\infty) \), moreover, if \( w^* < +\infty \), we have

\[
\lim_{w \to w^*+0} a(w) = a(w^*); \quad \lim_{w \to +\infty} a(w) = 0 \tag{1.5}
\]

where \( a(w^*) \) may or may not be finite. Thus, if \( a(w^*) \geq 1 \), the equation \( a(w) = 1 \) has a unique positive root \( w_0 > 1 \).

**Proposition 1.1.** For every memoryless channel with the property \( a(w^*) > 1 \) (in particular, for every memoryless channel with finite alphabet) there exists a positive constant \( b \) such that

\[
b_0 w_0^t \leq N(t) \leq w_0^t \quad (t \geq 0) \tag{1.6}
\]

where \( w_0 \) is the unique positive root of the equation \( a(w) = 1 \).

In particular, the channel capacity \( C \) exists and

\[
C = \log_2 w_0. \tag{1.7}
\]

Furthermore, if \( a(w^*) \leq 1 \), the capacity \( C \) still exists and it is equal to \( \log_2 w^* \).

**Proof.** If \( 0 \leq t < t^* \), \( E(t) \) consists of the void sequence only, thus \( N(t) = 1 \). If \( t \geq t^* \), the void sequence does not belong to \( E(t) \) and the number of sequences in \( E(t) \) beginning with the letter \( y_j \in Y \) is obviously \( N(t - t_j) \). Hence

\[
N(t) = \sum_{j \in J} N(t - t_j) \quad \text{if} \quad t \geq t^* \tag{1.8}
\]

and \( N(t) = 0 \) if \( t < 0 \), \( N(t) = 1 \) if \( 0 \leq t < t^* \).

We prove (1.6) by induction. Define the positive numbers \( b_k \) recursively by

\[
b_k w_0^t \leq N(t) \leq w_0^t \quad (0 \leq t < k t^*). \tag{1.9}
\]

and assume that

\[
b_k w_0^t \leq N(t) \leq w_0^t \quad (0 \leq t < k t^*). \tag{1.10}
\]

(1.10) clearly holds for \( k = 1 \); moreover, if it holds for some positive integer \( k \), for \( k t^* \leq t < (k + 1)t^* \) we have, in view of the assumption \( a(w_0) = \sum_{j \in J} w_0^{-t_j} = 1 \),

\[
b_{k+1} w_0^t \leq \sum_{j \leq i \leq k t^*} b_k w_0^{-t_j} \leq \sum_{j \in J} N(t - t_j) \leq \sum_{j \in J} w_0^{-t_j} = w_0^t. \tag{1.11}
\]
Thus, on account of (1.8), (1.10) holds for \( k + 1 \) if it holds for \( k \). Then by induction (1.10) holds for all \( k \) (observe that the sequence \( b_k \) is non-increasing).

To complete the proof of (1.6), we have only to show that \( b = \lim_{k \to \infty} b_k \) is positive. This is trivial if the alphabet \( Y \) is finite. Otherwise we have to show that the infinite product \( \prod_{k=1}^{\infty} (\sum_{t_j \leq k} w_i^{-t_j}) \) converges.

The convergence of this product is equivalent to the convergence of the series \( \sum_{k=1}^{\infty} \sum_{t_j > k} w_i^{-t_j} \) (according to the well-known lemma that an infinite product \( \prod_{k=1}^{\infty} (1 - a_k) \) (0 < \( a_k < 1 \) converges if and only if the series \( \sum_{k=1}^{\infty} a_k \) converges). The latter series, however, after changing the order of summations, is bounded from above by \( 1/t* \sum_{j \in J} t_j w_i^{-t_j} \); as, by the assumption \( a(w^*) > 1 \), the series (1.4) converges also for some \( w < w_0 \), the series

\[
-w_0 a'(w_0) = \sum_{j \in J} t_j w_0^{-t_j}
\]  

(1.12)
is convergent. The proof of (1.6) and (1.7) is complete.

To prove the last statement of Proposition 1.1, observe that in the case \( J = \{1, 2, \cdots \} \), \( a(w^*) \leq 1 \) one verifies by the same induction as above that

\[
N(t) \leq w^* t \quad \text{for all } t.
\]  

(1.13)

On the other hand, by the definition of \( w^* \), one may find to every \( \epsilon > 0 \) a positive integer \( M \) with \( \sum_{j=1}^{M} (w^* - \epsilon)^{-t_j} > 1 \); hence, for some \( w'_0 \) with \( w^* - \epsilon < w'_0 < w^* \) we have \( \sum_{j=1}^{M} w_i^{-t_j} = 1 \), implying (by what has already been proved) that the capacity of the memoryless channel with alphabet \( Y = \{y_j\}_{j \in J'}, J' = \{1, \cdots, M\} \) and symbol costs \( t_j \) (\( j \in J' \)) equals \( C' = \log_2 w'_0 \). Thus we arrive at

\[
\lim_{t \to \infty} \frac{\log_2 N(t)}{t} \geq \log_2 w'_0 \geq \log_2 (w^* - \epsilon).
\]  

(1.14)

As \( \epsilon > 0 \) has been arbitrary, (1.13) and (1.14) give rise to \( C = \log_2 w^* \). The proof of Proposition 1.1 is complete.

Let now \( X = \{x_i\}_{i \in I} \) be another (finite or countably infinite) alphabet, and let \( g \) be a mapping of \( X \) into \( Y^* \). For \( u = x_{i_1} \cdots x_{i_n} \in X^* \) let \( g(u) \) be the concatenation of the "code words" \( g(x_{i_k}) \) (\( k = 1, \cdots, n \)), i.e.

\[
g(x_{i_1} \cdots x_{i_n}) = g(x_{i_1}) \cdots g(x_{i_n}).
\]  

(1.15)

The mapping \( g \) is said to be a decodable code if \( u_1 \in X^*, u_2 \in X^*, u_1 \neq u_2 \) implies \( g(u_1) \neq g(u_2) \). We are going to give a simple information-
theoretic proof of the following well-known proposition (see, e.g., Krause (1962)), and we shall also indicate a possible simplification of the familiar combinatorial proof of this proposition.

**Proposition 1.2.** Let $P = \{p_i\}_{i \in I}$ be a probability distribution with entropy

$$H = H(P) = -\sum_{i \in I} p_i \log_2 p_i$$

(1.16)

and let $g$ be a decodable code assigning to $x_i \in X$ the code word $g(x_i) = y_{j_1(i)} \cdots y_{j_n(i)} \in Y^*$. Define

$$L = \sum_{i \in I} p_i l_i; \quad l_i = \sum_{k=1}^{n_i} t_{j_k(i)}$$

(1.17)

Then

$$L \geq \frac{H}{C_0}$$

(1.18)

where $C_0$ is the channel capacity defined by (1.3).

**Proof.** Let $\xi_1, \xi_2, \cdots$ be a sequence of independent random variables with range $X$ and common distribution $P(\xi_n = x_i) = p_i$ $(i \in I, n = 1, 2, \cdots)$. Let the random variable $X_n$ take on the value $l_i$ (see (1.17)) if $\xi_n = x_i$ and set

$$\tau_n = \sum_{k=1}^n \lambda_k.$$

(1.19)

As $g$ is a decodable code, under the condition $\tau_n = t$ the random sequence $\xi_1, \cdots, \xi_n$ can take on at most $N_0(t)$ different “values” $u \in X^*$. Thus the average conditional entropy of $\xi_1, \cdots, \xi_n$ given $\tau_n$ is upper-bounded by the expectation of $\log_2 N_0(\tau_n)$ and we have

$$nH(\xi_1) = H(\xi_1 \cdots \xi_n) \leq E \log_2 N_0(\tau_n) + H(\tau_n).$$

(1.20)

Here, in view of (1.3') and of $\tau_n \geq nt^*$ (see (1.1) and (1.19)) we have for any fixed $\epsilon > 0$

$$\log_2 N_0(\tau_n) \leq (1 + \epsilon)C_0\tau_n \quad \text{if} \quad n \geq n_0(\epsilon).$$

(1.21)

Suppose first that the $\xi$'s have only $m$ possible values, $x_1, \cdots, x_m$, say. Then, as $\tau_n$ is uniquely determined by the frequencies of the $x_i$'s $(1 \leq i \leq m)$ in the sequence $\xi_1, \cdots, \xi_n$, the number of possible values
of \( \tau_n \) is upperbounded by \((n + 1)^m\); thus \( H(\tau_n) \leq m \log_2 (n + 1) \) and

\[
\frac{1}{n} H(\tau_n) \to 0 \quad \text{as} \quad n \to \infty. \tag{1.22}
\]

Now, dividing (1.20) by \( n \), letting \( n \to \infty \) and taking into account (1.21), (1.22) and the identity \( E\tau_n = nE\lambda_1 = nL \), we arrive at

\[
H = H(\xi) \leq C_0 L. \tag{1.23}
\]

Thus (1.18) is valid if \( P = \{p_i\}_{i \in I} \) is a finite distribution; in the other case we may write \( P = \{p_1, p_2, \ldots\} \) and consider instead of \( P \) the finite distribution \( P^{(m)} = \{a_m p_1, \ldots, a_m p_m\}, a_m = (\sum_{i=1}^{m} p_i)^{-1} \). By what has already been proved, there holds

\[
L^{(m)} = a_m \sum_{i=1}^{m} p_i l_i \geq \frac{H^{(m)}}{C_0} = \frac{-\sum_{i=1}^{m} a_m p_i \log_2 a_m p_i}{C_0}. \tag{1.24}
\]

If \( m \to \infty \), we have \( a_m \to 1 \), \( L^{(m)} \to L \) and \( H^{(m)} \to H \), thus (1.18) holds for the general case as well.

\textbf{Remark.} Let, in particular, \( X = Y \) and let \( p_j = w_0^{-t_j}, j \in J = I \) (provided that \( w_0 \) satisfying \( \sum_{j \in J} w_0^{-t_j} = 1 \) does exist, i.e. provided that \( a(w^*) \geq 1 \)). Let the code \( g \) be the identity mapping. Then \( l_i = t_i \) and thus, using (1.7), we have

\[
CL = \log_2 w_0 \cdot \sum_{i \in I} w_0^{-t_i} \cdot t_i = -\sum_{i \in I} w_0^{-t_i} \cdot \log_2 w_0^{-t_i} = H.
\]

Comparing this with (1.18), we obtain \( C_0 \geq C \); the opposite inequality being obvious, we have arrived at \( C_0 = C \), i.e. the capacity in both senses is the same. This argument fails in the case \( a(w^*) < 1 \) but an approximation argument still suffices.

Proposition 1.2 is a particular case of Proposition 2.2 of Csiszár (1969) and its above proof is, essentially, the adaptation of the proof given there to the present simple case. Observe, too, that in the case \( t_j = 1 \) for all \( j \in J \) and \( J = \{1, \ldots, m\} \) proposition 1.2 reduces to the familiar simplest form of the "noiseless coding theorem" asserting

\[
L = \sum_{i \in I} p_i n_i \geq \frac{H}{\log_2 m}. \tag{1.25}
\]

In the author's mind, the above proof deserves attention even for this simplest case, and, compared with the familiar proof (see e.g. Feinstein
(1958), it has the merit of avoiding combinatorial arguments and using the properties of the amount of information only.

Let us make a comment also to the usual proof of the inequality \( L \geq H/C \), which starts by verifying first the inequality

\[
\sum 2^{-c_{li}} \leq 1,
\]

(1.26)

see Krause, 1962. The comment to be made consists in pointing out a simple direct way of verifying (1.26), using an idea of Karush (1961).

In fact, suppose first \( I = \{1, \ldots, m\} \) and let us denote the left hand side of (1.26) by \( A \). Let \( M \) be an arbitrary positive integer. Then

\[
A^M = \sum_{\sum M_i = M} \frac{M!}{M_1! \cdots M_m!} 2^{-c \sum_{i=1}^{m} M_i l_i} \leq \sum_{k=M}^{bM} B(M, k) 2^{-c (k-1) t^*}
\]

(1.27)

where

\[
B(M, k) = \sum_{\sum M_i = M} \frac{M!}{M_1! \cdots M_m!}
\]

(1.28)

and \( b \) is the smallest integer satisfying \( l_i \leq bt^* \) for each \( i = 1, \ldots, m \).

Then \( B(M, k) \) is just the number of sequences \( u = x_{i_1} \cdots x_{i_M} \in X^* \) with \( g(u) \in E(kt^*) \). Thus, in view of the decodability assumption and the definition of \( N(t) \)

\[
B(M, k) \leq N(kt^*) \quad k = M, M + 1, \ldots
\]

(1.29)

and, on account of (1.3), we arrive at

\[
B(M, k) \leq 2^{(c+\epsilon)kt^*} \quad k = M, M + 1, \ldots
\]

(1.30)

if \( M \geq M(\epsilon), \epsilon > 0 \) being arbitrary.

(1.27) and (1.30) gives rise to

\[
A^M \leq 2^{ct^* \cdot bM} \cdot 2^{\epsilon kt^*}
\]

(1.31)

and as \( \epsilon > 0 \) has been arbitrarily small, (1.31) can hold for every \( M \geq M(\epsilon) \) only if \( A \leq 1 \).

In the case \( I = \{1, 2, \ldots\} \) the above argument shows that
\[
\sum_{i=1}^{m} 2^{-c_i} \leq 1 \text{ for every positive integer } m, \text{ thus (1.26) has been proved for this case, as well.}
\]

2. A finite-state channel with alphabet \( Y = \{y_j\}_{j \in J} \) and set of states \( S = \{s_1, \ldots, s_r\} \) is defined, see Shannon and Weaver (1949), by specifying

(i) for each \( s_i \in S \) the set \( Y(s_i) = \{y_j\}_{j \in J(i)} \) of symbols transmissible by the channel at the state \( s_i \);

(ii) for each pair \( s_i, s_j \in S, y_j \in Y(s_i) \) the cost \( t_{ij} \geq 0 \) of transmission of the symbol \( y_j \) if the state is \( s_i \);

(iii) a function \( F(i, j) \) \( (i = 1, \ldots, r; j \in J(i)) \) specifying the new state \( s_{F(i, j)} \in S \) if the symbol \( y_j \) has been transmitted at the state \( s_i \).

A sequence \( u = y_{j_1} \cdots y_{j_n} \in Y^* \) is transmissible by the channel with initial state \( s_i \in S \) if

\[
j_k \in J(i_{k-1}) \quad k = 1, \ldots, n
\]

the \( i_k \)'s being defined recursively by

\[
i_k = F(i_{k-1}, j_k) \quad k = 1, \ldots, n, \quad i_0 = i; \quad (2.2)
\]

\( S_{i_k} \) is the state of the channel after having transmitted the \( k \)th symbol \( y_{j_k} \).

Let \( V_i \) denote the set of all sequences \( u \in Y^* \) which are transmissible by the channel with initial state \( s_i \). If \( t \geq 0 \), let \( E_i(t) \) denote the set of sequences \( u = y_{j_1} \cdots y_{j_n} \in V_i \) satisfying

\[
\sum_{k=1}^{n} y_{i_{k-1}j_k} \leq t, \quad \sum_{k=1}^{n} t_{i_{k-1}i_k} + t_{i_k}^* > t \quad (2.3)
\]

where

\[
t_{i_k}^* = \inf_{j \in J(i_k)} t_{ij} \quad i = 1, \ldots, r. \quad (2.4)
\]

We do not exclude the possibility of \( t_{i_k}^* = 0 \) for some \( i \), but we assume, as a part of the definition of a finite-state channel, that for every \( 1 \leq i \leq r \) and every \( y_{j_1} \cdots y_{j_n} \in V_i \) with \( n \geq r \) there exist at least one \( k \leq n \) for which \( t_{i_k}^* > 0 \).

The channel is called indecomposable if for each pair of states \( s_i, s_k \) there exists a sequence \( y_{j_1} \cdots y_{j_n} \in V_i \) for which \( i_n = k \) (see (2.1) and (2.2)); here, of course, one may assume \( n \leq r \).

Let \( N_i(t) \) denote the number of sequences in \( E_i(t) \) if \( t \geq 0 \) and set \( N_i(t) = 0 \) if \( t < 0 \). Define
\[ C_i = \lim_{t \to \infty} \frac{1}{t} \log_2 N_i(t). \]  

(2.5)

If \( C_i \) exists for every \( 1 \leq i \leq r \) and its value does not depend on \( i \), this common value is called the capacity \( C \) of the finite-state channel under consideration.

To this definition similar remarks are due as to (1.3). A rigorous proof of Proposition 2.1 below, for the finite alphabet case with integer symbol costs, has been given by Ljubič (1962). We are going to give a simpler proof valid for arbitrary symbol costs as well.

Let \( \alpha(w) \) denote the \( r \times r \) matrix

\[ \alpha(w) = (a_{ik}(w)), \quad a_{ik}(w) = \sum_{j \in J_k(i)} w^{-i_j} \]  

(2.6)

where

\[ J_k(i) = \{ j : j \in J(i), F(i,j) = k \} \]  

(2.7)

**Proposition 2.1.** Let us be given an indecomposable finite-state channel with finite alphabet. Then there exist positive numbers \( 0 < b < B \) such that

\[ bw_0^t \leq N_i(t) \leq Bw_0^t, \quad i = 1, \ldots, r \]  

(2.8)

where \( w_0 \) is defined as the (unique) positive number \( w \) for which the greatest positive eigenvalue of the matrix \( \alpha(w) \) equals one. This \( w_0 \) is the greatest positive root of the equation

\[ \text{Det} \left[ \sum_{j \in J_k(i)} w^{-i_j} \delta_{ik} \right] = 0 \]  

(2.9)

where \( \delta_{ik} \) is Kronecker's delta.

In particular, the capacity of the channel exists and it is equal to

\[ C = \log_2 w_0. \]  

(2.10)

The capacity formula (2.10) remains valid for channels with infinite alphabet, as well, with the only modification that if \( w_0 \) defined above does not exist, it should be replaced by the infimum \( w^* \) of the positive numbers \( w \) for which all entries of the matrix \( \alpha(w) \) are finite.

**Proof.** If \( 0 \leq t < t_i^* \), the set \( E_i(t) \) consists of the void sequence only; thus \( N_i(t) = 1 \) if \( 0 \leq t < t_i^* \). If \( t \geq t_i^* \), the void sequence does not belong to \( E_i(t) \) and the number of sequences in \( E_i(t) \) beginning with \( y_j (j \in J(i)) \) is equal to \( N_{F(i,j)}(t - t_{ij}) \). Hence, in view of (2.7),

\[ N_i(t) = \sum_{k=1}^{r} \sum_{j \in J_k(i)} N_k(t - t_{ij}) \quad \text{if} \quad t \geq t_i^* \quad (1 \leq i \leq r) \]  

(2.11)
As $\alpha$ is a matrix with nonnegative elements, its greatest positive eigenvalue $\lambda(w)$ is equal to the least upper bound of the set of positive $\lambda$'s satisfying $\sum_{k=1}^{r} a_{ik}(w) a_k \geq \lambda a_i$, $i = 1, \ldots, r$, for some $r$-tuple of nonnegative numbers $a_1, \ldots, a_r$, not all equal to zero. Moreover, as the matrix $\alpha(w)$ is indecomposable (this is obviously equivalent to the indecomposability of the channel), the components of the eigenvector of $\alpha(w)$ belonging to its greatest positive eigenvalue are positive.

The above representation of $\lambda(w)$ implies that $\lambda(w)$ is a strictly decreasing, continuous function of $w$, with $\lambda(1) \geq 1$ and $\lim_{w \to \infty} \lambda(w) = 0$. Thus there exists a unique positive number $w_0 \geq 1$ with $\lambda(w_0) = 1$. Let the components of the eigenvector of $\alpha(w_0)$ belonging to the eigenvalue $\lambda(w_0) = 1$ be denoted by $a_i, i = 1, \ldots, r$. Then $a_i > 0$, $i = 1, \ldots, r$ and

$$
\sum_{k=1}^{r} \sum_{j \in J_k(i)} a_k w_0^{-1+} = \sum_{k=1}^{r} a_{ik}(w_0) a_k = a_i \quad (1 \leq i \leq r). \quad (2.12)
$$

Let now $a$ and $A$ be positive numbers such that in the interval $0 \leq t < T = \max_{1 \leq i \leq r, j \in J(i)} r_{ij}$

$$
aa_i w_0^t \leq N_i(t) \leq Aa_i w_0^t, \quad i = 1, \ldots, r. \quad (2.13)
$$

As the numbers $a_i$ are all positive and $T$ is finite (we are considering the finite alphabet case), such positive numbers $a < A$ surely exist.

Then one verifies in the same way as in the proof of Proposition 1.1 that (2.13) holds for every $t \geq 0$. In fact, if

$$
t^* = \min_{1 \leq i \leq r, j \in J(i)} t_{ij} > 0 \quad (2.14)
$$

and (2.13) is valid for $0 \leq t < T + nt^*$ then for $T + nt^* \leq t < T + (n + 1)t^*$ we have $0 \leq t - t_{ij} < T + nt^* \quad (1 \leq i \leq r, j \in J(i))$ and

$$
\sum_{k=1}^{r} \sum_{j \in J_k(r)} a_{ik} w_0^{t - t_{ij}} \leq \sum_{k=1}^{r} \sum_{j \in J_k(r)} Aa_k w_0^{t - t_{ij}} = Aa_i w_0^t; \quad (2.15)
$$

this means, on account of (2.11), that (2.13) holds for $0 \leq t < T + (n + 1)t^*$ as well.

Concerning the simple properties of matrices with nonnegative elements used below, see e.g. Gantmacher (1959).
If $t_{ij} = 0$ for some pair $(i, j)$ i.e. if $t^* = 0$, the above induction breaks down. However, in view of the assumption formulated in connection with (2.4), one has for every finite-state channel
\[ t^{**} = \min_{1 \leq i \leq r} \sum_{k=1}^{r} t_{ik-1} > 0. \] (2.16)

Then the above argument can be modified by letting $t^{**}$ and $rT$ play the role of $t^*$ and $T$, respectively, and referring to the $r'$th iterates of the systems of equations (2.11) and (2.12) (instead of (2.11) and (2.12) themselves).

Thus, for any indecomposable finite-state channel with finite alphabet, the inequalities (2.13) hold for every $t \geq 0$, proving (2.8) and (2.10).

Obviously, $w_0$ is a positive root of equation (2.9). Moreover, as the greatest positive eigenvalue $\lambda(w)$ of $a(w)$ is a strictly decreasing function of $w$, in the case $w > w_0$ the number 1 cannot be an eigenvalue of $a(w)$ thus $w$ cannot be a root of (2.9); i.e., $w_0$ is the greatest positive root of equation (2.9).

If the channel alphabet $Y$ is infinite and $w^*$ is the greatest lower bound of the positive numbers $w$ for which all entries of the matrix $a(w)$ are finite, the greatest positive eigenvalue $\lambda(w)$ of $a(w)$ is still a strictly decreasing continuous function of $w$ for $w > w^*$. Thus, if $\lambda(w) \geq 1$ for some $w > w^*$ then there exists a unique $w_0 > w^*$ with $\lambda(w) = 1$ and $w_0$ is the greatest positive root of equation (2.9). Now, if $\lambda(w) \leq 1$ for some $w \geq w^*$, one verifies in the same way as in the finite alphabet case that
\[ N_i(t) \leq A a_i w^i \quad (1 \leq i \leq r) \] (2.17)
for every real number $t$, where the $a_i$'s are the components of the eigenvector of $a(w)$ belonging to the eigenvalue $\lambda(w)$ (observe that from the point of view of the proof of the second part of (2.13), the role of $T$ can be played by $\max_{1 \leq i \leq r} t_i^*$, as well). Furthermore, if for a positive number $w$ either $w > w^*$, $\lambda(w) \geq 1$ or $w < w^*$, one can find, for any $\epsilon > 0$, a positive integer $M$ such that restricting the alphabet $Y = \{y_1, y_2, \cdots \}$ to $Y' = \{y_1, \cdots, y_M\}$, the number $w_0'$—defined as $w_0$ for the restricted channel—satisfies $w_0' > w - \epsilon$. Thus, in the same way as in the case of Proposition 1.1, we arrive at the proof of the last statement of Proposition 2.1, as well.

Remark. For the infinite-alphabet case we have shown that if $\lambda(w) \geq 1$ for some $w > w^*$ then there exists a unique $w_0 > w^*$ with $\lambda(w_0) = 1$. 
and for every $t \geq 0$

$$b(w_0 - \epsilon)^t \leq N_i(t) \leq Bw_0^t, \quad i = 1, \cdots, r \quad (2.18)$$

where $b$ and $B$ are positive constants, $b$ depending on $\epsilon$. It would be desirable to prove that in (2.18) $\epsilon$ may be omitted, just as in the case of memoryless channels, see Proposition 1.1; to this end, however, the above argument seems to need some nontrivial refinement.

Finally, let us describe an encoding procedure for finite-state channels. Let us be given a finite-state channel with finite alphabet and let $X = \{x_i\}_{i \in I}$ ($I = \{1, \cdots, m\}$ or $I = \{1, 2, \cdots\}$) be another alphabet. Let $P = \{p_i\}_{i \in I}$ be a given probability distribution on $X$ with finite entropy

$$H = H(P) = -\sum_{i \in I} p_i \log_2 p_i; \quad (2.19)$$

without any loss of generality we assume $p_1 \geq p_2 \geq p_3 \geq \cdots$. We make correspond to each $x_i \in X$ a code word $g_k(x_i) = y_{j_1} \cdots y_{j_n}$ (depending on $k$, where $1 \leq k \leq r$) in the following way. Set

$$\alpha_i = \sum_{i=1}^{i-1} p_i \quad (i \in I) \quad (2.20)$$

Let us subdivide the unit interval into $|J(i_0)|$ disjoint subintervals $I_{j_1}$, $j_1 \in J(i_0)$ of length $(1/a_i)\Delta x_{F(i_0, i_1)}w_0^{-1}f_0 I_{j_1}$ where the numbers $a_k$ and $w_0$ are the same as in the proof of Proposition 2.1. In the next step, let us subdivide each $I_{j_1}$ ($j_1 \in J(i_0)$) into $|J(i_1)|$ disjoint subintervals $I_{j_1j_2}$ ($j_2 \in J(i_1)$) of length $(1/a_{j_1})\Delta x_{F(i_1, j_2)}w_0^{-1}f_0 I_{j_1j_2}$, etc.; in the $n$th step let us subdivide each $I_{j_1\cdots j_{n-1}}$ into $|J(i_{n-1})|$ disjoint subintervals $I_{j_1\cdots j_n}$ ($j_n \in J(i_{n-1})$) of length $1/a_{i_0}a_{F(i_{n-1}, j_0)}w_0^{-1}f_0 I_{j_1\cdots j_n}$. Now we set, see (2.20), $g_k(x_i) = y_{j_1} \cdots y_{j_n}$ if $\alpha_i \in I_{j_1\cdots j_n}$ and neither $\alpha_{i+1}$ nor $\alpha_{i-1}$ belongs to $I_{j_1\cdots j_n}$ while either $\alpha_{i+1}$ or $\alpha_{i-1}$ (or both) belongs to $I_{j_1\cdots j_{n-1}}$.

**Proposition 2.2.** The encoding $x_i \rightarrow g_k(x_i)$ described above has the prefix property; the code words $g_k(x_i)$ are transmissible by the channel if the initial state is $i_0 = k$ and the average code cost $L = \sum_{i \in I} p_i d_i$ satisfies

$$L < \frac{H + B}{C} + T \quad (2.21)$$

Here by interval we mean a left semi-closed interval. $|J(i_0)|$ denotes the number of elements of the set $J(i_0)$.
\[
l_i = \sum_{k=1}^{n} t_{i_{k-1}i_k} \quad \text{if} \quad g_k(x_i) = y_{i_1}\cdots y_{i_n}. \tag{2.22}
\]

B and T are constants depending on the channel only and C = \(\log_2 w_0\) is the channel capacity.

**Proof.** The prefix property of the encoding \(x_i \rightarrow g_k(x_i)\) and the assertion that \(g_k(x_i)\) is transmissible if the initial state is \(k\) i.e. that \(g_k(x_i) \in V_k\) follow from the very construction of the \(g_k(x_i)\)'s, cf. (2.1) and (2.2). It is also obvious from the construction that if \(g_k(x_i) = y_{i_1} \cdots y_{i_n}\) then the length of \(I_{i_1}\cdots i_{n-1}\) is greater than

\[
\min (\alpha_i - \alpha_{i-1}, \alpha_{i+1} - \alpha_i) = p_i.
\]

This means

\[
\frac{1}{a_{i_0}} a_{i_{n-1}} w_0^{-1} \sum_{i=1}^{n-1} t_{i_{i-1}i} > p_i. \tag{2.23}
\]

Write \(B' = \max_{1 \leq i \leq r, 1 \leq k \leq r} a_i/a_k\) and \(T = \max_{1 \leq i \leq r, j \in J(i)} t_{ij}\); then (2.22) and (2.23) imply

\[
(l_i - T) \log_2 w_0 < -\log_2 p_i + \log_2 B' \quad (i \in I); \tag{2.24}
\]

hence, on account of (2.10) and (2.19) we obtain (2.21), with \(B = \log_2 B'\).

Let us remark that the above encoding procedure can be considered as a generalization for finite-state channels of the encoding suggested by Krause (1962) for memoryless channels which, in turn, has been a generalization to the case of different symbol costs of the familiar Shannon-Fano code.

If we are given an arbitrary stationary source \(\mathcal{X}\) with alphabet \(X\) and with entropy rate \(H(\mathcal{X}) = H < \infty\), on the basis of Proposition 2.2 one arrives in a well-known way at the result that the output of the source can be encoded in such a way—encoding sufficiently large blocks—that the average code cost per source symbol be arbitrarily close to \(H\). Concerning the problems arising in connection with the “noiseless coding theorem” we refer to Csiszár-Katona-Tusnády (1969) and Csiszár (1969) where a general treatment of the subject is given.

**RECEIVED:** December 17, 1968
REFERENCES


