# ON ACHIEVING CHANNEL CAPACITY IN FINITE-MEMORY CHANNELS

BY

LEO BREIMAN[1]

## Introduction

For finite-memory channels as defined in [1] it has been an open question since Hinčin's paper [2] as to whether the ergodic capacity equals the stationary capacity.[2] That is, using $R(p)$ to denote the rate achieved on the channel by input measure $p$, is it true that

(a) $$\sup_p \{R(p); p \text{ stationary}\} = \sup_p \{R(p); p \text{ ergodic}\}?$$

We show not only equality but, denoting the common value by $C$, that there is at least one ergodic measure $p$ such that

(b) $$C = R(p).$$

The method used depends essentially on the following theorem.

THEOREM 1. *An upper-semicontinuous (U.S.C.) bounded linear functional defined on a convex compact subset of a linear locally convex separated topological space assumes its supremum on at least one of the extreme points of the set.*

Theorem 1 is a simple consequence of the Kreĭn-Mil'man theorem (see, for example [3]), and its proof is relegated to an appendix since theorems of the same nature can be found in the literature. Its relevance is that, looked at in the right way, the set of stationary input measures is a convex compact subset of a linear topological space whose extreme points are the ergodic input measures. This fact, along with other definitions, we develop in the first section. Secondly, we show that for a general channel, not necessarily of finite memory, the rate $R(p)$ is a linear function of the input measure. In the third section we show that in the correct sense $R(p)$ is an U.S.C. functional for channels of finite memory.

*Note added after completion of the paper.* There have been a number of recent papers relevant to this one, and a few remarks on these would not be inappropriate. Hinčin's definition in [2] of finite-memory channels seems to be insufficiently restrictive to establish his indicated results. Takano in [4] imposed severer conditions, using what he termed $m$-finite memory $m$-dependent channels in order to derive the results of [2]. This latter class of

Received January 31, 1959.

[1] This research was started at the University of California, Berkeley, with the support of the Office of Naval Research, and was finished while the author was a National Science Foundation fellow.

[2] This statement is no longer true. See the remarks at the end of the introduction.

channels we refer to, following the terminology of Feinstein [1], as finite-memory channels, and they form the concern of the present paper. Tsaregradsky [5] proved (a) using Hinčin's definition of finite memory. However, his proof is not easy to follow, and it is not clear whether implicit use is not made of other restrictions. Feinstein in [6] published an elegant and elementary proof of (a) and other significant results for finite-memory channels as defined here. Both of the above proofs have the same plan: A new definition of capacity is introduced leading to a number $C_0$ such that $C_0$ is obviously larger than the stationary capacity. Then for any $\varepsilon > 0$, an ergodic input $p$ is constructed such that $R(p) > C_0 - \varepsilon$. The definition of $C_0$ is the same in both proofs and was also introduced by Wolfowitz [7]. The major part of the present work was finished before the above results became known to us and obviously proceeds along much different lines. Although our approach is not elementary, it gives, as a by-product, the proof of (b), which we doubt can be established by the more elementary methods used in [5] and [6].

## Definitions, notations, and assorted facts

We define very briefly a channel (see [1], [4], or [8] for a fuller definition). Suppose we are given an alphabet $D$, that is, a finite set of symbols $(d_1, \cdots, d_t)$. We denote by $\Omega(D)$ the set of all doubly infinite sequences $(\cdots, z_{-1}, z_0, z_1, \cdots)$ such that each coordinate takes values in $D$, by $\mathfrak{F}(D)$ the field of all finite-dimensional cylinder sets, and by $\mathfrak{B}(D)$ the Borel field generated by $\mathfrak{F}(D)$. Any cylinder set in $\mathfrak{F}(D)$ of the type $\{k^{\text{th}}$ coordinate is $z_k, \cdots, n^{\text{th}}$ coordinate is $z_n\}$ we refer to as the finite message $(z_k, \cdots, z_n)$ of length $n - k + 1$.

A channel consists of an output alphabet $B$ and an input alphabet $A$ together with a set of conditional probabilities

$$P(y_k, \cdots, y_n \mid \cdots, x_{-1}, x_0, x_1, \cdots)$$

defined for every finite message $(y_k, \cdots, y_n)$ in $\mathfrak{F}(B)$ and for every sequence $(\cdots, x_{-1}, x_0, x_1, \cdots)$ in $\Omega(A)$. We consider the general channel to be stationary and nonanticipatory.

For the channel to have finite memory, there must be an integer $m$ such that

(i) for all $n, k$, $n \geqq k$

$$P(y_k, \cdots, y_n \mid \cdots x_n) = P(y_k, \cdots, y_n \mid x_{k-m}, \cdots, x_n),$$

(ii) for all $k, i, j, n$, $n \geqq j \geqq i \geqq k$ and $j - i > m$

$$P(y_k, \cdots, y_i, y_j, \cdots, y_n \mid \cdots, x_n)$$
$$= P(y_k, \cdots, y_i \mid x_{k-m}, \cdots, x_i)P(y_j, \cdots, y_n \mid x_{j-m}, \cdots, x_n).$$

The smallest integer $m$ for which (i) and (ii) hold is called the memory length.

The set of all finite measures on $\mathcal{B}(A)$ forms a linear space $\mathcal{G}$. We topologize $\mathcal{G}$ so that $\mu_\nu \to \mu$ if and only if $\mu_\nu(x_k, \cdots, x_n) \to \mu(x_k, \cdots, x_n)$ for every finite message $(x_k, \cdots, x_n)$. In this topology we have

THEOREM 2. *$\mathcal{G}$ is a linear locally convex separated topological space, and the set $\mathcal{P}$ of probability measures on $\mathcal{B}(A)$ is a compact subset of $\mathcal{G}$.*

As with Theorem 1, we give a brief sketch of the proof of Theorem 2 in the appendix. The set of stationary probability measures $\mathcal{S}$, that is, the set of all probability measures invariant under the shift operation, is easily seen to be closed, therefore compact, and obviously convex. We assert

THEOREM 3. *The set of extreme points $\mathcal{E}$ of $\mathcal{S}$ is exactly the set of ergodic stationary measures on $\mathcal{B}(A)$.*

*Proof.* Let $\mu$ be stationary and ergodic, and assume $\mu = \alpha\mu_1 + \beta\mu_2$, $\alpha + \beta = 1$, $\alpha, \beta > 0$, $\mu_1, \mu_2 \in \mathcal{S}$. Suppose $\mu_1$ is not ergodic, and let $S$ be an invariant Borel set such that $0 < \mu_1(S) < 1$. Then $0 < \mu(S) < 1$, which shows that both $\mu_1$ and $\mu_2$ must be ergodic. A consequence of the ergodic theorem is that any two distinct ergodic measures on $\mathcal{B}(A)$ are $\perp$. Let $S$ now be an invariant Borel set such that $\mu_1(S) = 1$, $\mu_2(S) = 0$; we have the contradiction $0 < \mu(S) < 1$. Now let $\nu$ be an extreme point of $\mathcal{S}$, and suppose that $\nu$ is not ergodic. Then there is an invariant set $S$ in $\mathcal{B}(A)$ such that $0 < \nu(S) < 1$. We define two stationary probability measures $\nu_1$, $\nu_2$ as follows. If $E \in \mathcal{B}(A)$, then

$$\nu_1(E) = \nu(S \cap E)/\nu(S), \qquad \nu_2(E) = \nu(S^c \cap E)/\nu(S^c).$$

Hence $\nu(E) = \nu(S)\nu_1(E) + \nu(S^c)\nu_2(E)$ so that $\nu$ must be ergodic.

We proceed to define the rate of a channel which is a functional $R(p)$ on $\mathcal{S}$. For any alphabet $D$, with $q$ a stationary probability measure on $\mathcal{B}(D)$, we form the functions of $q$

$$H_n(q) = -(1/n) \sum_z q(z) \log q(z),$$

where $z$ is any message $n$ long beginning at the first coordinate, and the summation is over all such messages. McMillan [8] has shown that $H_n(q) \geqq H_{n+1}(q)$, and $H(q) = \lim_n H_n(q)$ is called the entropy of the process. If we have a channel and a stationary probability $p$ on $\mathcal{B}(A)$, then we use the conditional probabilities of the channel in the time-honored fashion to put a stationary probability $p''$ on the Borel field of the sequence space $\cdots, (x_{-1}, y_{-1}), (x_0, y_0), \cdots, x_k \in A, y_k \in B$. The marginal of $p''$ induces a stationary probability $p'$ on $\mathcal{B}(B)$. The expression

$$R(p) = H(p) + H(p') - H(p'')$$

is the rate produced by the source $p$.

## 1. The additive property of entropy and rate

THEOREM 4. *Let* $p, q \in \mathcal{S}$, $\alpha + \beta = 1$, $\alpha > 0$, $\beta > 0$. *Then*

$$H(\alpha p + \beta q) = \alpha H(p) + \beta H(q),$$

$$R(\alpha p + \beta q) = \alpha R(p) + \beta R(q).$$

*Proof.* Let $x$ denote messages of length $n$.

$$\begin{aligned} H_n(\alpha p + \beta p) &= -(1/n) \sum_x [\alpha p(x) + \beta q(x)] \log [\alpha p(x) + \beta q(x)] \\ &= -(1/n) \sum_x \alpha p(x) \log p(x) - (1/n) \sum_x \beta q(x) \log q(x) \\ &\quad - (1/n) \sum_x \alpha p(x) \log [\alpha + \beta\, q(x)/p(x)] \\ &\quad - (1/n) \sum_x \beta q(x) \log [\beta + \alpha\, p(x)/q(x)]. \end{aligned}$$

We use

$$\log \alpha \leqq \log [\alpha + \beta\, q(x)/p(x)] \leqq \log \alpha + (\beta/\alpha)(q(x)/p(x))$$

to deduce

$$\begin{aligned} (1/n)\, \alpha \log \alpha \leqq (1/n) \sum_x \alpha p(x) \log [\alpha + \beta\, q(x)/p(x)] \\ \leqq (1/n)\, \alpha \log \alpha + (1/n)(\beta/\alpha) \end{aligned}$$

and conclude that the third term on the right above goes to zero. We treat the fourth term similarly and conclude that $H(p)$ is additive. To check the additivity of $R(p)$ we have only to note that $p'$ and $p''$ are linearly dependent on $p$.

## 2. $R(p)$ is an U.S.C. functional on $\mathcal{S}$ for finite-memory channels

THEOREM 5. *$H(p')$ is an U.S.C. functional on $\mathcal{S}$ for channels of finite memory length $m$.*

*Proof.* $H(p') = \lim_n H_n(p')$ where $H_{n+1}(p') \leqq H_n(p')$. Take any sequence $\{p_N\} \subset \mathcal{S}$ such that $p_N \to p$, let $y$ be any message $k - m$ long in $\mathfrak{F}(B)$ beginning at coordinate $m + 1$, and let $x$ denote messages of length $k$ in $\mathfrak{F}(A)$ beginning at coordinate one. Then $p'_N(y) = \sum_x P(y \mid x) p_N(x)$ so that $p'_N(y) \to p'(y)$ as $p_N \to p$. Thus

$$-(1/n) \sum_y p'_N(y) \log p'_N(y) \to -(1/n) \sum_y p'(y) \log p'(y)$$

as $p_N \to p$, so that $H_n(p')$ is a continuous function on $\mathcal{S}$. Thus $H(p')$, being the limit of a decreasing sequence of continuous functions, is U.S.C.

To deal with the rest of $R(p)$ we introduce functions $L_n(p)$ on $\mathcal{S}$ as follows: Let $y$ be any message $n - m$ long in $\mathfrak{F}(B)$ beginning at $m + 1$, and $x$ any message $n$ long in $\mathfrak{F}(A)$ beginning at one; then define

$$L_n(p) = (1/n) \sum_{x,y} P(y \mid x) p(x) \log P(y \mid x).$$

PROPOSITION 1. *For a channel with memory length $m$,*

$$L_n(p) \to H(p) - H(p'').$$

*Proof.*

$$L_n(p) = (1/n) \sum_{x,y} P(y \mid x)p(x) \log P(y \mid x)p(x) + H_n(p)$$
$$= (1/n) \sum_{x,y} p''(y, x) \log p''(y, x) + H_n(p).$$

Let $\bar{y}$ stand for any message $n$ long in $\mathfrak{F}(B)$ that starts at one and coincides with $y$ from $m + 1$ on, and $\underline{x}$ any message $n - m$ long in $\mathfrak{F}(A)$ that starts at $m + 1$ and coincides thereafter with $x$. Then $\log p''(\bar{y}, x) \leqq \log p''(y, x)$, so

$$(1/n) \sum_{x,\bar{y}} p''(\bar{y}, x) \log p''(\bar{y}, x) \leqq (1/n) \sum_{x,\bar{y}} p''(\bar{y}, x) \log p''(y, x)$$
$$= (1/n) \sum_{x,y} p''(y, x) \log p''(y, x).$$

The left-hand side above is $H_n(p)$; passing to the limit gives

$$H(p) - H(p'') \leqq \liminf L_n(p).$$

On the other hand, $\log p''(y, x) \geqq \log p''(y, \underline{x})$, yielding

$$(1/n) \sum_{x,y} p''(y, x) \log p''(y, x) \geqq (1/n) \sum_{y,x} p''(y, x) \log p''(y, \underline{x})$$
$$= (1/n) \sum_{y,\underline{x}} p''(y, \underline{x}) \log p''(y, \underline{x}).$$

The right-hand side is equal to $-(n - m/n)H_n(p'')$, and the limit gives $H(p) - H(p'') \geqq \limsup L_n(p)$. Now we prove

THEOREM 6. *For a channel with finite memory length $m$, $H(p) - H(p'')$ is an U.S.C. functional on* S.

*Proof.* Let $n = k(m + l)$, $k > 1$, let $x$, $y$ be as before, let $u_i$, $v_i$, $w_i$ be the parts of the $x$ and $y$ messages as in these diagrams

$$\overbrace{x_1, \cdots, x_{m+l}}^{w_1}, \cdots, \overbrace{\cdots, x_{2(m+l)}}^{w_2}, \cdots, \overbrace{x_{(k-1)(m+l)+1}, \cdots, x_{k(m+l)}}^{w_k},$$

$$\overbrace{y_{m+1}, \cdots, y_{m+l}}^{u_1}, \overbrace{\cdots, y_{2m+l}}^{v_1}, \overbrace{\cdots, y_{2(m+l)}}^{u_2}, \overbrace{\cdots, y_{2m+3l}}^{v_2}, \cdots,$$

$$\overbrace{y_{km+(k-1)l}, \cdots, y_{k(m+l)}}^{v_{k-1} \quad u_k}.$$

We write

$$L_n(p) = (1/n) \sum_{u,v,w} P(u_1, v_1, \cdots, v_{k-1}, u_k \mid w_1, \cdots, w_k)p(w_1, \cdots, w_k)$$
$$\cdot \log P(u_1, v_1, \cdots, u_k \mid w_1, \cdots, w_k),$$

but

$$\log P(u_1, v_1, \cdots, v_{k-1}, u_k \mid w_1, \cdots, w_k)$$
$$\leqq \log P(u_1, \cdots, u_k \mid w_1, \cdots, w_k) = \sum_{i=1}^{k} \log P(u_i \mid w_i).$$

Therefore,

$$L_n(p) \leqq \frac{1}{k(l+m)} \sum_{i=1}^{k} \sum_{u_1,w_i} P(u_i \mid w_i)p(w_2) \log P(u_i \mid w_i)$$

$$= \frac{1}{l+m} \sum_{u_1,w_1} P(u_1 \mid w_1)p(w_1) \log P(u_1 \mid w_1) = L_{l+m}(p).$$

Now $L_n(p)$ depends only on finite-dimensional distributions, and just as in Theorem 5, we easily prove that $L_n(p)$ is a continuous function of $p$. By the inequality $L_{k(l+m)}(p) \leqq L_{l+m}(p)$ we may choose a subsequence of the $L_n$ converging monotonically downward to $H(p) - H(p'')$, which proves the theorem.

Since the sum of two U.S.C. functions is again U.S.C., we have shown that for finite-memory channels $R(p)$ is U.S.C. on $\mathcal{S}$. When we combine this fact with the previous theorems, the result announced in the introduction is assembled.

## Appendix

*Proof of Theorem* 1. Let $\mathfrak{N}$ be a compact, convex subset of a linear locally convex separated topological space, and $f(x)$ a bounded U.S.C. functional on $\mathfrak{N}$. Since $f$ is U.S.C. and $\mathfrak{N}$ compact, $f$ assumes its supremum on $\mathfrak{N}$. Let $\mathfrak{M}$ be the subset of $\mathfrak{N}$ on which $f$ equals its supremum; then by U.S.C. $\mathfrak{M}$ is closed, and by the linearity at $f$, $\mathfrak{M}$ is convex. By the Kreĭn-Mil'man theorem $\mathfrak{M}$ must have at least one extreme point. Let $x$ be such an extreme point; then $x$ must be extreme in $\mathfrak{N}$. For if $x = \alpha x_1 + \beta x_2$, $\alpha + \beta = 1$, $\alpha, \beta > 0$, $x_1, x_2 \in \mathfrak{N}$, then $f(x) = \alpha f(x_1) + \beta f(x_2)$ which implies that both $x_1$ and $x_2$ are in $\mathfrak{M}$, and proves the theorem.

*Proof of Theorem* 2. We consider $\Omega(A)$ as $\prod_{-\infty}^{+\infty} A_i$, where each $A_i$ is a copy of $A$. In the product topology, by Tychonoff's theorem, $\Omega(A)$ is compact. The space $\mathcal{G}$ is the adjoint of the space $\mathcal{C}$ of all continuous functions on $\Omega(A)$. In the weak dual topology on $\mathcal{G}$, i.e., that topology such that

$$\mu_N \to \mu \quad \Leftrightarrow \quad \int f \, d\mu_n \to \int f \, d\mu$$

for every $f \in \mathcal{C}$, the unit sphere in $\mathcal{G}$ is compact. (See, for example, [9].) The set of probability measures on $\mathcal{B}(A)$ form a closed, and therefore compact, subset of the unit sphere. It remains to show that the weak dual topology is equivalent to convergence on finite-dimensional cylinder sets. One way is quick; if $\mu_N \to \mu$ in the weak dual topology, then for any finite-dimensional cylinder set $S$, we have that $I_S(x)$ is continuous, and hence $\mu_N(s) \to \mu(s)$. Going the other way is accomplished by using the elementary result that every continuous function on $\Omega(A)$ can be uniformly approximated by a finite linear combination of indicators of finite-dimensional cylinder sets, which follows from the compactness of $\Omega(A)$.

It is a pleasure to acknowledge the pleasant and illuminating discussions we have had with A. Feinstein touching on the subject treated above.

*Note added in proof.* We have been informed that K. R. Parthasaraty of the Indian Statistical Institute has recently used similar methods to derive the result (a) above.

## References

1. A. Feinstein, *Foundations of information theory*, New York, 1958.
2. A. Hinčin, *On the fundamental theorems of information theory*, Uspehi Mat. Nauk (N.S.), vol. 11 (1956), no. 1 (67), pp. 17–75 (in Russian).
3. N. Bourbaki, *Espaces vectoriels topologiques*, Éléments de Mathématique, Livre V, Actualités Scientifiques et Industrielles, no. 1189, Paris, 1953, Chapter II, p. 84.
4. K. Takano, *On the basic theorems of information theory*, Ann. Inst. Statist. Math., Tokyo, vol. 9 (1958), pp. 53–77.
5. I. P. Tsaregradsky, *On the capacity of a stationary channel with finite memory*, Teor. Veroyatnost. i Primenen, vol. 3 (1958), pp. 84–96 (in Russian).
6. A. Feinstein, *On the coding theorem and its converse for finite-memory channels*, Technical Report No. 41 (June 20, 1958), Applied Mathematics and Statistics Laboratory, Stanford University (to be published in Communication and Control).
7. J. Wolfowitz, *The maximum achievable length of an error correcting code*, Illinois J. Math., vol. 2 (1958), pp. 454–458.
8. B. McMillan, *The basic theorems of information theory*, Ann. Math. Statistics, vol. 24 (1953), pp. 196–219.
9. L. H. Loomis, *Generalized harmonic analysis*, New York, 1956.

University of California
Berkeley, California
Paris, France