The Annals of Mathematical Statistics
1969, Vol. 40, No. 6, 2156–2177

ON THE OPTIMUM RATE OF TRANSMITTING INFORMATION[1]

BY J. H. B. KEMPERMAN

University of Rochester

1. Summary. The present paper is partly expository and does not assume any previous knowledge of information theory or coding theory. It is meant to be the first in a series of papers on coding theory for noisy channels, this series replacing the report [12] which was widely circulated. A few of the results were reported in [11], [27] and [28].

In Sections 2 and 3 we present certain refinements and generalizations of known methods of Shannon, Fano, Feinstein and Gallager. A discussion of some other methods may be found in the surveys by Kotz [15] and Wolfowitz [28] such as those of Khintchine [14], McMillan [18] and Wolfowitz [25], [26].

Section 4 contains a number of applications. Most of this section is devoted to a certain memoryless channel with additive noise. Some of the proofs have been collected in Section 5. Finally, Section 6 describes some new results on the relative entropy $H(\mu_1 \mid \mu_2)$, of one measure μ_1 relative to another measure μ_2, and its relation with the total variation $\|\mu_1 - \mu_2\|$.

2. Terminology. In the sequel N we will denote a positive integer and ϵ a more or less fixed constant, $0 < \epsilon < 1$. All logarithms are to the base e. We further take $0 \cdot \infty = 0$. By ϕ we denote the continuous convex function on $[0, +\infty)$ defined by $\phi(z) = z \log z$, $\phi(0) = 0$. Most measures employed are probability measures. Further $\mu \ll \nu$ will denote that the measure μ is absolutely continuous with respect to the measure ν.

2.1. By a *channel* we mean a non-empty collection of probability measures on a fixed measurable space $Y = (Y, \mathfrak{F}_Y)$. Thus $S = \{P_x(B); x \varepsilon X, B \varepsilon \mathfrak{F}_Y\}$. The index set X is called the *input alphabet* of the channel S while Y is called the *output alphabet* of S. If X is finite then S is said to be *semicontinuous*. If both X and Y are finite sets then S is said to be a *discrete* channel. We often write $P_x(B) = P(B \mid x)$. One may interpret S as a noisy channel where the sender transmits some $x \varepsilon X$. In that case the receiver will obtain a random signal η taking values in Y such that $\Pr(\eta \varepsilon B \mid x) = P(B \mid x)$.

2.2. The direct *product* $S^{(n)} = S_1 \times \cdots \times S_n$ of a sequence of channels

$$S_m = \{P_m(B \mid x); x \varepsilon X_m, B \subset Y_m\}, \quad (m = 1, \cdots, n), \tag{2.1}$$

is defined as the channel

$$S^{(n)} = \{P^{(n)}(B \mid u); u \varepsilon X^{(n)}, B \subset Y^{(n)}\},$$

Received 26 May 1969.

[1] This work was supported by the National Science Foundation. A preliminary version of the present paper appeared in [13]. The research was further supported by the Office of Naval Research.

where $X^{(n)} = X_1 \times \cdots \times X_n$, $Y^{(n)} = Y_1 \times \cdots \times Y_n$. Further, if $u = (x_1, \cdots, x_n) \varepsilon X^{(n)}$ then

$$P^{(n)}(\cdot \mid u) = P_1(\cdot \mid x_1) \times \cdots \times P_n(\cdot \mid x_n).$$

Thus, if a sender transmits the *word* $u = (x_1, \cdots, x_n)$ over this product channel $S^{(n)}$ the received channel will be a random word $\eta = (\eta_1, \cdots, \eta_n)$ belonging to $Y^{(n)}$, that is, the mth received letter η_m takes values in Y_m. Moreover, the η_m are independent random variables such that Pr $(\eta_m \varepsilon B_m \mid u) = P_m(B_m \mid x_m)$ for each $B_m \varepsilon \mathfrak{F}_{Y_m}$. In particular, the distribution of η_m depends only on the mth letter x_m transmitted. This explains why $S^{(n)}$ is also called a *memoryless channel.*

The memoryless channel $S^{(n)}$ is said to be *stationary* if all the components S_m of $S^{(n)}$ are copies of one and the same channel S. In that case we write $S^{(n)} = S^n$.

2.3. From now on S will be a fixed channel as in **2.1** (possibly itself a product of other channels). Further, L denotes a subset of the input alphabet X of S.

2.4. An ϵ-code for S is defined as any sequence $\{(x^{(i)}, D^{(i)}); i = 1, \cdots, N\}$ such that $x^{(i)} \varepsilon X$, while the $D^{(i)}$ are *disjoint* measurable subsets of Y satisfying

$$P(D^{(i)} \mid x^{(i)}) \geqq 1 - \epsilon \quad \text{for each} \quad i = 1, \cdots, N. \tag{2.2}$$

Instead of an ϵ-code one also speaks of a code having a maximal error $\leqq \epsilon$, (when the sender restricts himself to the symbols $x^{(1)}, \cdots, x^{(N)}$, while the receiver concludes from $\eta \varepsilon D^{(i)}$ that the sender transmitted $x^{(i)}$).

If (2.2) is replaced by the weaker condition

$$N^{-1} \textstyle\sum_{i=1}^{N} P(D^{(i)} \mid x^{(i)}) \geqq 1 - \epsilon \tag{2.3}$$

we speak of a code for S having an average error $\leqq \epsilon$.

2.5. Let $L \subset X$, $0 < \epsilon < 1$. Let $N_L(S, \epsilon)$ denote the supremum of the set of integers N for which there exists an ϵ-code of length N as above for which $x^{(i)} \varepsilon L$, $(i = 1, \cdots, N)$. Similarly $\bar{N}_L(S, \epsilon)$ for codes with an average error $\leqq \epsilon$. As is easily seen,

$$(1 - c))\bar{N}_L(S, c\epsilon) \leqq N_L(S, \epsilon) \leqq \bar{N}_L(S, \epsilon) \quad \text{for each} \quad 0 < c < 1. \tag{2.4}$$

When $L = X$ we simply write $N(S, \epsilon)$ and $\bar{N}(S, \epsilon)$, respectively. We shall be especially interested in the asymptotic behavior of $N(S^n, \epsilon)$ when n is large. However, if $N(S, \delta) = +\infty$ for one number $0 < \delta < 1$ then one easily sees ([1] page 16) that, for each $0 < \epsilon < 1$, one has $N(S^n, \epsilon) = +\infty$ as soon as $n \geqq n_0(\epsilon)$.

For this reason we are more interested in channels S such that $N(S, \epsilon) < \infty$ for each $0 < \epsilon < 1$. For such channels one can always find a probability measure ν on Y such that

$$P_x(\cdot) = P(\cdot \mid x) \ll \nu \quad \text{for each} \quad x \varepsilon X. \tag{2.5}$$

A simple proof may be found in **5.1.**

2.6. Let ν be a fixed probability measure on Y satisfying (2.5). Let

$$f_x(y) = f(x, y) = dP_x/d\nu \qquad (x \varepsilon X, y \varepsilon Y), \tag{2.6}$$

denote the corresponding Radon-Nikodym derivatives (densities). Thus, $f_x(y) = f(x, y)$ is non-negative and measurable in y. Moreover,

$$P(D \mid x) = \int_D f_x(y)\nu(dy) = \int_D f(x, y)\nu(dy),$$

for $x \varepsilon X$, $D \varepsilon \mathfrak{F}_Y$. Note that $\int f_x \, d\nu = 1$. For each fixed $x \varepsilon X$, the function f_x is unique up to a change in a set $E \subset Y$ with $\nu(E) = 0$.

2.7. *A measurable function on a probability space may always be regarded as a random variable.* It will be convenient to regard the measurable function

$$J(x \mid \nu) = \log\,(dP_x/d\nu) = \log f(x, y) \tag{2.7}$$

as a random variable, employing $(Y, \mathfrak{F}_Y, P_x)$ as the underlying probability space. We have $J(x \mid \nu) > -\infty$ with probability 1 since the set $\{y : f_x(y) = 0\}$ has a P_x-measure 0. The corresponding distribution function is given by

$$F(\theta \mid x, \nu) = \Pr\,(J(x \mid \nu) \leqq \theta) = \int_{f_x \leqq e^\theta} dP_x = \int_{f \leqq e^\theta} f(x, y)\nu(dy). \tag{2.8}$$

Its expectation value is given by

$$E\{J(x \mid \nu)\} = \int \log f_x(y) P_x(dy) = \int \phi(f_x(y))\nu(dy) = H(P_x \mid \nu). \tag{2.9}$$

Here, $H(P_x \mid \nu) \geqq 0$ is precisely the entropy of P_x relative to ν. A more detailed discussion of such entropies may be found in Section 6.

2.8. In the sequel we want to avoid the explicit assumption (2.5), (though some results would be trivial without it). Then (2.6) is to be interpreted as the density of the absolutely continuous component of P_x, everything relative to ν. We shall further take $f(x, y)$ and $J(x \mid \nu)$ as $+\infty$ on some set D_x with $\nu(D_x) = 0$ and $P_x(D_x)$ maximal. Thus $J(x \mid \nu) = +\infty$ with a positive probability (positive P_x-measure) precisely when P_x is not absolutely continuous with respect to ν.

It is often easier to treat this case as follows. Select any σ-finite measure λ on Y such that both P_x and ν are absolutely continuous with respect to λ. Put $g(x, y) = dP_x/d\lambda$, $h(y) = d\nu/d\lambda$. Define $J(x \mid \nu)$ as the measurable function

$$J(x \mid \nu) = \log\,[g(x, y)/h(y)] \tag{2.10}$$

on the probability space $(Y, \mathfrak{F}_Y, P_x)$. It is independent of the particular choice of λ, (up to a change in a set of P_x-measure 0). One has $J(x \mid \nu) = +\infty$ at each point $y \varepsilon Y$ where $g(x, y) > 0$ and $h(y) = 0$. Further,

$$E\{J(x \mid \nu)\} = \int \log\,[g(x, y)/h(y)]P_x(dy) = H(P_x \mid \nu). \tag{2.11}$$

3. Upper and lower bounds.

Let ν be a fixed probability measure on Y (not necessarily satisfying (2.5)). The following result has several interesting consequences. In some sense, it amounts to an exact formula for $\bar{N}_L(S, \epsilon)$.

LEMMA 3.1. *Let N be a positive integer. In order that $N \leqq \bar{N}_L(S, \epsilon)$ it is necessary and sufficient that one can find N elements $x^{(i)} \varepsilon L$ $(i = 1, \cdots, N)$, not necessarily distinct, such that*

$$\int \{\max_{i=1,\cdots,N} f(x^{(i)}, y)\}\nu(dy) \geqq (1 - \epsilon)N. \tag{3.1}$$

The latter is equivalent to

$$\int \{\textstyle\sum_{i=1}^{N} f(x^{(i)}, y) - \max_{i=1,\cdots,N} f(x^{(i)}, y)\}\nu(dy) \leqq \epsilon N. \tag{3.2}$$

Here ν is any probability measure on Y such that the $P(\cdot \mid x^{(i)})$ $(i = 1, \cdots, N)$ are absolutely continuous relative to ν.

Note that the left hand sides of (3.1) and (3.2) do not depend on the precise choice of the measure ν. The proof of Lemma 3.1 is given in **5.2.**

Lemma 3.2. *Let θ and ρ be real constants such that*

$$F(\theta \mid x, \nu) \geqq \epsilon + e^{-\rho} \quad \text{for each} \quad x \,\varepsilon\, L. \tag{3.3}$$

Then

$$N_L(S, \epsilon) \leqq \bar{N}_L(S, \epsilon) \leqq e^{\theta+\rho}. \tag{3.4}$$

3.1. The proof of Lemma 3.2 is given in **5.3.** It is essential that ν be a probability measure, that is, $\nu(Y) = 1$. In the case that (2.5) fails to hold one must interpret f and F as in **2.8.**

The validity of (3.3) would normally depend on the actual choice of ν. In principle, one would like to choose the probability measure ν in such a way that (3.3) holds with $\theta + \rho$ as small as possible. A result of the type found in Lemma 3.2 was stated and used for the fisst time by the author (see [12] page 7 and [27]).

3.2. The following result is useful in conjunction with (3.1). For each choice of the non-negative numbers $f_1, \cdots, f_N$ we have the inequality

$$\begin{aligned} &N^{-1}\log N \max(f_1, \cdots, f_N) \\ &\qquad \leqq N^{-1}\textstyle\sum_{i=1}^{N}\phi(f_i) - \phi(N^{-1}\sum_{i=1}^{N} f_i) + N^{-1}\log 2\sum_{i=1}^{N} f_i. \end{aligned} \tag{3.5}$$

The proof of (3.5) is given in **5.4.**

Let $N \leqq \bar{N}_L(S\epsilon)$; thus there exist elements $x^{(i)} \,\varepsilon\, L$ $(i = 1, \cdots, N)$ satisfying (3.1). Put $f_i = f(x^{(i)}, \cdot)$ and $g = N^{-1}\sum_{i=1}^{N} f_i$, thus, $\int g\, d\nu = 1$. Multiplying (3.1) by $N^{-1}\log N$ and applying (3.5), we have the following inequality of Fano [4].

$$\begin{aligned} (1 - \epsilon)\log N &\leqq \log 2 + \int \{N^{-1}\textstyle\sum_{i=1}^{N}\phi(f_i) - \phi(g)\}\nu(dy) \\ &= \log 2 + \textstyle\sum_{i=1}^{N} N^{-1}\int \log(f_i/g)P(dy \mid x^{(i)}) \\ &= \log 2 + \int H(P_x \mid \nu^{\Pi})\Pi(dx). \end{aligned} \tag{3.6}$$

Here, Π denotes the probability measure on X of finite support

$$\{x_1, \cdots, x_N\} \subset L$$

defined by $\Pi(A) = N^{-1}$ [no. of $j = 1, \cdots, N$ with $x_j \,\varepsilon\, A$], (A running through the σ-field of all subsets of X). Further, ν^{Π} denotes the measure on Y defined by

$$\nu^{\Pi}(B) = \int P(B \mid x)\Pi(dx), \tag{3.7}$$

where B runs through $\mathfrak{F}_Y$. Finally, see (6.2), $H(P_x \mid \nu^{\Pi})$ is precisely the entropy of P_x relative to ν^{Π}.

3.3. Let $Q = Q(X)$ denote the collection of all probability measures on X of finite support. Similarly $Q(L)$ when $L \subset X$. To each $\Pi \varepsilon Q$ we associate the probability measure ν^{Π} on Y defined by (3.7). Motivated by (3.6), let us further ntro duce, for each $\Pi \varepsilon Q$,

$$C(\Pi) = \int H(P_x \mid \nu^{\Pi})\Pi(dx) \tag{3.8}$$

and

$$C_L = \sup_{\Pi \varepsilon Q(L)} C(\Pi), \qquad (L \subset X). \tag{3.9}$$

When $L = X$ we simply write $C_L = C$ and call C the *Shannon capacity* of the channel S. With the above notations, (3.6) may be written as

$$(1 - \epsilon) \log \bar{N}_L(S, \epsilon) \leqq C_L + \log 2. \tag{3.10}$$

It is customary (see [28]) to interpret $C(\Pi)$ as the expected value of a random variable $J(\Pi)$ as follows. Given $\Pi \varepsilon Q$, let ν^{Π} be defined as in (3.7), thus, $P(\cdot \mid x) \ll \nu^{\Pi}$ for each x in the support of Π. Put $f(x, y) = dP(\cdot \mid x)/d\nu^{\Pi}$. By (3.8),

$$C(\Pi) = \int \int \log f(x, y) P(dy \mid x)\Pi(dx) = E\{J(\Pi)\}. \tag{3.11}$$

Here, $J(\Pi)$ corresponds to the measurable function $\log f(x, y)$ on the probability space $(X \times Y, \mathfrak{F}_X \times \mathfrak{F}_Y, Q^{\Pi})$, where

$$Q^{\Pi}(A \times B) = \int_A P(B \mid x)\Pi(dx), \qquad (A \varepsilon \mathfrak{F}_X, B \varepsilon \mathfrak{F}_Y). \tag{3.12}$$

In the present case, we take $\mathfrak{F}_X$ as the σ-field of all subsets of X; however, the above formulae also make sense in certain other situations where Π has an infinite support. The distribution function of $J(\Pi)$ is given by

$$F(\theta \mid \Pi) = \int_{f \leqq e^{\theta}} P(dy \mid x)\Pi(dx) = \int F(\theta \mid x, \nu^{\Pi})\Pi(dx). \tag{3.13}$$

The above formulae are complementary to those in **2.7.**

3.4. The following result is essentially from Feinstein [6], [7] page 46; see also [2] page 1232 and [28] page 91. For the benefit of the reader its proof is reproduced in **5.5.**

LEMMA 3.3. *We have for each $\Pi \varepsilon Q$ and each real number θ that*

$$\bar{N}(S, \epsilon) \geqq N(S, \epsilon) \geqq [\epsilon - \Pr(J(\Pi) \leqq \theta)]e^{\theta} \tag{3.14}$$

and

$$N_L(S, \epsilon) \geqq [\int_L \{\epsilon - F(\theta \mid x, \nu^{\Pi})\}\Pi(dx)]e^{\theta}. \tag{3.15}$$

Consequently,

$$N_L(S, \epsilon) \geqq \Pi(L)e^{\theta - \rho} \tag{3.16}$$

as soon as θ and ρ are real constants with

$$F(\theta \mid x, \nu^{\Pi}) \leqq \epsilon - e^{-\rho} \quad \textit{for each} \quad x \varepsilon L. \tag{3.17}$$

REMARK 3.1. Feinstein's Lemma 3.3 carries over to more general probability measures Π. More precisely, let $\mathfrak{F}_X$ denote a σ-field of subsets of X and Π a probability measure on $\mathfrak{F}_X$; here, $\mathfrak{F}_X$ may vary with Π; usually, it is more convenient to replace $\mathfrak{F}_X$ by its completion relative to Π.

We shall assume (a) that $P(B \mid x)$ is an $\mathfrak{F}_X$-measurable function of x for each fixed $B \varepsilon \mathfrak{F}_Y$. Then (3.7) defines a measure ν^{Π} on $\mathfrak{F}_Y$. We shall further assume (b) that the Radon-Nikodym derivative $f(x, y)$ of (the absolutely continuous component of) $P(\cdot \mid x)$ relative to ν^{Π} can be chosen as a jointly measurable function on $X \times Y$, when supplied with the σ-field $\mathfrak{F}_X \times \mathfrak{F}_Y$. Then Q^{Π} and $J(\Pi)$ can be defined in a way completely analogous to the special case in **3.3.** However, comparing **2.8**, one must allow the possibility that $J(\Pi) = +\infty$ with a positive probability, precisely when $P(\cdot \mid x) \ll \nu^{\Pi}$ does not hold for almost all [Π] elements $x \varepsilon X$.

Let Q^* denote the collection of all such measures Π, more precisely, the collection of all pairs $(\mathfrak{F}_X, \Pi)$ having the above properties. As follows from the proof in **5.5**, Lemma 3.3 remains valid when Q is replaced by Q^*.

There is one minor difficulty, namely, the set $L \subset X$ may not be measurable relative to $\mathfrak{F}_X$. This can be taken care of by replacing in (3.16) the quantity $\Pi(L)$ by the corresponding outer measure. Similarly for the integral in (3.15).

3.5. A different lower bound on $\bar{N}_L(S, \epsilon)$, especially useful when ϵ is small, can be obtained from Lemma 3.1 by a form of the method of random codes due to Shannon [21], compare **3.9.**

3.6. *Definitions.* Consider a pair $(\mathfrak{F}_X, \Pi)$ of the following type. First, $\mathfrak{F}_X$ is a σ-field of subsets of X while Π is a probability measure on $\mathfrak{F}_X$. A greater degree of generality may be achieved in the sequel by assuming that $\mathfrak{F}_X$ is complete relative to Π.

Next, we assume that a probability measure ν on $Y = (Y, \mathfrak{F}_Y)$ can be found such that $P_x \ll \nu$ holds for almost [Π] all $x \varepsilon X$ and such that $f(x, y) = dP_x/d\nu$ can be chosen as a jointly measurable function of x and y relative to $\mathfrak{F}_X \times \mathfrak{F}_Y$. The precise choice of ν is unimportant. The collection of all such pairs $(\mathfrak{F}_X, \Pi)$ will be denoted as Q^{**}. One has $Q \subset Q^{**} \subset Q^*$, provided $\mathfrak{F}_x$ is always assumed to be complete.

Given such a pair and an associated measurable density $f(x, y) = dP(\cdot \mid x)/d\nu$, define

$$g_t(y) = [\int f(x, y)^t \Pi(dx)]^{1/t}, \tag{3.18}$$

$$G_t = G_t(\Pi) = \int g_t(y)\nu(dy), \tag{3.19}$$

and

$$C_t = C_t(\Pi) = -t(1 - t)^{-1} \log G_t(\Pi). \tag{3.20}$$

Here, $0 < t < 1$. It is easily seen that G_t and C_t do not depend on the particular choice of ν. Further, $g_t(y) \leqq \int f(x, y)\Pi(dx)$ while $\int f(x, y)\nu(dy) = 1$, hence, $G_t \leqq 1$ and $C_t \geqq 0$.

The following lemma is a straightforward generalization of a result of Gallager [9] which in turn is related to results of Shannon [21] and Fano [5], Chapter 9.

LEMMA 3.4. *Let* $\Pi = (\mathfrak{F}_X, \Pi)$ *be a fixed pair in* Q^{**} *and let* $C_t = C_t(\Pi)$ *be defined as above. Then we have for each number* $\frac{1}{2} \leqq t < 1$ *that*

$$\bar{N}(S, \epsilon) \geqq [\exp\{C_t - t(1-t)^{-1}\log \epsilon^{-1}\}], \tag{3.21}$$

where $[z]$ *denotes the integer part of* z.

3.7. By (2.4) with $c = \frac{1}{2}$, one has $N(S, \epsilon) \geqq \frac{1}{2}\bar{N}(S, \frac{1}{2}\epsilon)$ thus (3.21) also yields lower bounds for $N(S, \epsilon)$. Another way of formulating (3.21) would be to fix $\frac{1}{2} \leqq t < 1$ and the positive integer $N < \exp\{C_t\}$. The assertion then is that there exists a code of length N for S having an average error $\leqq \epsilon$, where ϵ is defined by

$$\log N + t(1-t)^{-1}\log \epsilon^{-1} = C_t(\Pi). \tag{3.22}$$

Clearly, one would like to choose Π such that $C_t(\Pi)$ is as large as possible.

3.8. A proof of the following lemma may be found in **5.6.** This lemma is given in a form more general than needed (in the proof of Lemma 3.4), since it seems to be of independent interest.

LEMMA 3.5. *Let* $\frac{1}{2} \leqq t < 1$ *be a fixed number and let* $Z_1, \cdots, Z_N$ *denote non-negative random variables satisfying*

$$E\{Z_i^t Z_j^t\} \leqq E\{Z_i^t\}E\{Z_j^t\} \quad \text{whenever} \quad i \neq j. \tag{3.23}$$

Then

$$W = \textstyle\sum_{i=1}^N Z_i - \max_{i=1,\cdots,N} Z_i \tag{3.24}$$

satisfies

$$E\{W\} \leqq (\textstyle\sum_{i=1}^N E\{Z_i^t\})^{t^{-1}}. \tag{3.25}$$

3.9. *Proof of Lemma* 3.4. The following proof is closely related to Gallager's [9] proof. Let N be a fixed positive integer not exceeding the right hand side of (3.21). By (3.20), this is equivalent to $N^{t^{-1}}G_t \leqq \epsilon N$.

We must prove that $\bar{N}(S, \epsilon) \geqq N$. By Lemma 3.1, it suffices to prove the existence of N elements $x^{(i)} \varepsilon X$ $(i = 1, \cdots, N)$, not necessarily distinct, and satisfying (3.2). Sufficient for the latter is that

$$\textstyle\int \{\sum_{i=1}^N f(x^{(i)}, y) - \max_{i=1,\cdots,N} f(x^{(i)}, y)\}\nu(dy) \leqq N^{t^{-1}}G_t. \tag{3.26}$$

It suffices to show that (3.26) holds on the average. We shall average the left hand side of (3.26), (regarded as a function of $u = (x^{(1)}, \cdots, x^{(N)}) \varepsilon X^N$), with respect to the product measure $\Pi \times \cdots \times \Pi = \Pi^N$ on $X^N = (X, \mathfrak{F}_X)^N$. By Fubini, the two integrals can be interchanged, hence, it suffices to show that $\int \alpha(y)\nu(dy) \leqq N^{t^{-1}}G_t$, where

$$\alpha(y) = \textstyle\int \cdots \int \{\sum_{i=1}^N f(x^{(i)}, y) - \max_i f(x^{(i)}, y)\}\Pi(dx^{(1)}) \cdots \Pi(dx^{(N)}).$$

For each fixed y, we may interpret the measurable function $f(x^{(i)}, y)$ on the probability space $(X, \mathfrak{F}_X, \Pi)^N$ as a random variable $Z_i(y)$. Doing so, $\alpha(y)$ may be rewritten as

$$\alpha(y) = E\{\textstyle\sum_{i=1}^{N} Z_i(y) - \max_i Z_i(y)\}.$$

Since the $Z_i(y)$ are obviously independent, it follows from Lemma 3.5 that

$$\alpha(y) \leqq (NE\{Z_1(y)^t\})^{t^{-1}} = (N \textstyle\int f(x, y)^t \Pi(dx))^{t^{-1}}.$$

Using (3.18), (3.19) this implies $\int \alpha(y)\nu(dy) \leqq N^{t^{-1}}G_t$. This completes the proof of Lemma 3.4.

4. Memoryless channels. In the present paper we shall restrict ourselves to a few special applications of the auxiliary results presented in Section 3.

4.1. Let $S^{(n)}$ be a memoryless channel as in **2.2.** Let $Q^{(n)} = Q(X^{(n)})$ denote the collection of all measures of finite support on $X^{(n)} = X_1 \times \cdots \times X_n$. Just as in **3.3**, there corresponds to each $\Pi^{(n)} \varepsilon Q^{(n)}$ a random variable $J(\Pi^{(n)})$ having its mean equal to $C(\Pi^{(n)})$.

Now consider the special case of a product measure $\Pi^{(n)} = \Pi_1 \times \cdots \times \Pi_n$, where $\Pi_m \varepsilon Q_m = Q(X_m)$. It is easily seen that then $\nu^{\Pi^{(n)}}$ is a product measure on $Y^{(n)}$ just as $P(\cdot \mid u)$. It follows, see (3.13), that $J(\Pi^{(n)})$ is a sum of *n independent* random variables, the mth random variable being distributed as the random variable $J(\Pi_m)$ relative to S_m. In particular, taking expectations, we conclude that $C(\Pi_1 \times \cdots \times \Pi_n) = C(\Pi_1) + \cdots + C(\Pi_n)$. Without further assumptions not much can be said about the precise distribution of $J(\Pi^{(n)})$, as would be needed in applying (3.14).

4.2. For the sake of completeness, let us first present a known application, see [2] page 1233. Suppose that the above memoryless channel is stationary, $S^{(n)} = S^n$, and take all the components Π_m equal to the same $\Pi \varepsilon Q$. Then we obtain a random variable $J(\Pi^n)$ which is the sum of n independently and *identically distributed* random variables each having a mean $C(\Pi)$. It follows from the law of large numbers that

$$\lim_{n\to\infty} \Pr\,(J(\Pi^n) \leqq n\theta) = 0 \quad \text{when} \quad \theta < C(\Pi).$$

We can now conclude from (3.14) that

$$\liminf n^{-1} \log N(S^n, \epsilon) \geqq C(\Pi), \tag{4.1}$$

for each $\Pi \varepsilon Q = Q(X)$, (as n tends to infinity). Hence,

$$\liminf n^{-1} \log N(S^n, \epsilon) \geqq C, \tag{4.2}$$

where C denotes the Shannon capacity of S defined by (3.9).

It is not hard to show, see [28] page 102, that the Shannon capacity of S^n is equal to nC. Hence, by (3.10),

$$\limsup n^{-1} \log N(S^n, \epsilon) = \limsup n^{-1} \log \bar{N}(S^n, \epsilon) \leqq C/(1 - \epsilon); \tag{4.3}$$

(here, the equality sign follows easily from (2.4), at least for all but denumerably many values ϵ). Comparing (4.2) and (4.3), we see that $N(S^n, \epsilon) \approx e^{nC}$, at least for ϵ small but fixed. A result of this type was brilliantly conjectured by Shannon

[20] in 1948. Many rigorous proofs and extensions have since been given by Shannon, Fano, Feinstein, McMillan, Khintchine, Wolfowitz and many others, see [15] and [28].

4.3. In view of Remark 3.1 we have that (4.1) is true not only when $\Pi \varepsilon Q$ but also when $\Pi = (\mathfrak{F}_X, \Pi) \varepsilon Q^*$. Therefore, (4.2) even holds with C replaced by $C^* = \sup \{C(\Pi): \Pi \varepsilon Q^*\}$. Clearly, $C^* \geqq C$. On the other hand, using (4.3) we have $C^* \leqq C/(1 - \epsilon)$ for each $\epsilon > 0$, consequently, $C^* = C$.

In particular, if $C < \infty$ then we have for each $\Pi \varepsilon Q^*$ that $C(\Pi) < \infty$ implying that $\Pr (J(\Pi) = +\infty) = 0$, and hence that

$$P(\cdot \mid x) \ll \nu^{\Pi} \quad \text{for almost} \quad [\Pi] \quad \text{all} \quad x \varepsilon X.$$

It should be possible to prove this in a more elementary fashion. Note that the channel S itself is completely arbitrary.

4.4. *Additive noise.* Let G denote a fixed compact group, not necessarily commutative. The group operation in G will be taken as addition. Let μ denote the Haar measure on G, normalized such that $\mu(G) = 1$. Integrals with respect to μ are often written as $\int h(z)\mu(dz) = \int h(z)\, dz$. Further $\mathfrak{B}$ will denote the σ-field of Borel subsets of G.

We shall be interested in the special channel

$$S = \{P_x(B) = \eta(-x + B);\, x \varepsilon G,\, B \varepsilon \mathfrak{B}\}. \tag{4.4}$$

More precisely, we take both X and Y equal to G, $\mathfrak{F}_Y$ as $\mathfrak{B}$, and $P(B \mid x)$ as $\eta(-x + B)$, where η denotes a fixed regular probability measure on G. For this channel S, if the sender transmits the symbol $x \varepsilon G$ then the receiver receives the signal $x + W$, where W is a random variable with distribution $\Pr(W \varepsilon B) = \eta(B)$. Since this distribution does not depend on x one may speak of additive noise.

As pointed out in **2.5**, we are mainly interested in cases where one can find a probability measure ν such that $P(\cdot \mid x) \ll \nu$ for all x. In the present case this implies that $\eta \ll \mu$ (for a proof, see **5.7**). Put $f = d\eta/d\mu$, thus

$$\eta(B) = \int_B f(z)\mu(dz) = \int_B f(z)\, dz \quad \text{for all} \quad B \varepsilon \mathfrak{B}.$$

Clearly, any non-negative measurable function f, with $\int f\, dz = 1$, can arise in this manner. It is precisely the probability density function (relative to μ) of the above random variable W. Moreover,

$$f(x, y) = dP(\cdot \mid x)/d\mu = f(-x + y). \tag{4.5}$$

4.5. Let us now apply some of the results in Section 3, taking every time Π and ν as the Haar measure μ on G; by (3.17), we have also $\nu^{\Pi} = \mu$. It follows from (2.18), (3.13) and (4.5) that the distribution function $F(\theta \mid x, \nu)$ of $J(x \mid \nu)$, the distribution function $F(\theta \mid x, \nu^{\Pi})$ of $J(x \mid \nu^{\Pi})$ and the distribution function $F(\theta \mid \Pi)$ of $J(\Pi)$ *are all equal to one and the same function* $F(\theta)$ (independent of x). Namely,

$$F(\theta) = \int_{\log f(y) \leqq \theta} f(y)\, dy. \tag{4.6}$$

The corresponding expectation value

$$C(\Pi) = \int \{\log f(y)\} f(y)\, dy = \int_{-\infty}^{+\infty} \theta\, dF(\theta) \tag{4.7}$$

might be equal to $+\infty$. *One always has*

$$\int_{-\infty}^{+\infty} e^{-\theta}\, dF(\theta) \leqq 1, \tag{4.8}$$

since

$$\int e^{-\theta}\, dF(\theta) = \int e^{-\log f(y)} f(y)\, dy = \int_{f(y)>0} dy \leqq 1.$$

A proof of the following result may be found in **5.8**.

Lemma 4.1. *Every distribution function* $F(\theta)$ *on* $(-\infty, +\infty)$ *and satisfying* (4.8) *can arise in this manner.*

More precisely, given such a distribution function F *and letting*

$$g(v) = \int_{v+0}^{+\infty} w^{-1}\, d_w F(\log w), \tag{4.9}$$

and

$$f(s) = \inf \{v > 0 : g(v) \leqq s\}, \qquad (0 < s < 1), \tag{4.10}$$

we have the relation (4.6) *relative to the group* $G = [0, 1)$ *of the reals modulo* 1.

Applying Lemma 3.2 and Lemma 3.3, one obtains:

Lemma 4.2. *Let* $\theta_1 < \theta_2$ *be real numbers such that*

$$F(\theta_1) < \epsilon < F(\theta_2). \tag{4.11}$$

Then

$$e^{\theta_1}(\epsilon - F(\theta_1)) \leqq N(S, \epsilon) \leqq \bar{N}(S, \epsilon) \leqq e^{\theta_2}/(F(\theta_2) - \epsilon). \tag{4.12}$$

Further, using (3.18)–(3.20) and (4.5), we see that

$$C_t = -(1 - t)^{-1} \log \int f(y)^t\, dy = -\tau^{-1} \log \int e^{-\tau\theta}\, dF(\theta), \tag{4.13}$$

where $\tau = 1 - t$. Hence, Lemma 3.4 yields the following.

Lemma 4.3. *We have for each number* $0 < \tau \leqq \frac{1}{2}$ *that*

$$\bar{N}(S, \epsilon) \geqq [\exp -\tau^{-1} \{\log \int e^{-\tau\theta}\, dF(\theta) + (1 - \tau) \log \epsilon^{-1}\}]. \tag{4.14}$$

Here, $[z]$ *denotes the integral part of* z.

4.6. Let us say that the channel S described in **4.4** is of *normal type* (C, σ) if the associated function F is of the form

$$F(\theta) = \Phi(\sigma^{-1}(\theta - C)), \tag{4.15}$$

with Φ as the standardized normal distribution function

$$\Phi(z) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{z} e^{-\frac{1}{2}\zeta^2}\, d\zeta.$$

Here, $\sigma > 0$ and $C \geqq 0$. If F is concentrated at a single point C then S will be said to be of normal type $(C, 0)$. This happens precisely when f takes on only the two values 0 and e^C. In other words, when there is a set $E \in \mathcal{B}$ with $\mu(E) = e^{-C}$ such that $P_x(B) = e^C \mu((-x + B) \cap E)$ holds for all $x \in G$, $B \in \mathcal{B}$.

If S is of normal type (C, σ) then

$$\int e^{-\tau\theta}\, dF(\theta) = \exp\,(-C\tau + \tfrac{1}{2}\sigma^2\tau^2). \tag{4.16}$$

Hence, (4.8) would be equivalent to

$$\sigma \leqq (2C)^{\frac{1}{2}}. \tag{4.17}$$

In fact, by Lemma 4.1 the normal type (C, σ) *is possible if and only if* $C \geqq 0$, $\sigma \geqq 0$ in such a way that (4.17) holds.

4.7. For many (but not all) channels S it happens that the corresponding stationary memoryless channel S^n with n large behaves "nearly" as a channel of normal type (C, σ) with C of magnitude n and σ of magnitude $n^{\frac{1}{2}}$.

Thus, let us study $\bar{N}(S, \epsilon)$ for a channel of normal type (C, σ) with σ (and hence C) rather large. Let us first discuss the case that ϵ is very small, thus, the quantity γ defined by

$$\epsilon = e^{-\frac{1}{2}\gamma^2}, \qquad (\gamma > 0), \tag{4.18}$$

is very large. Analogous and much more detailed results were established by Shannon, Gallager and Berlekamp [22] for the case of a discrete stationary memoryless channel.

Lemma 4.4. *Let S be of normal type (C, σ). Then*

$$\begin{aligned} \bar{N}(S, \epsilon) &\geqq [\exp\{C - \tfrac{1}{2}\gamma(2\sigma - \gamma)\}] \quad \textit{when} \quad \gamma \leqq \tfrac{1}{2}\sigma, \\ &\geqq [\exp\{C - \tfrac{1}{4}\sigma^2 - \tfrac{1}{2}\gamma^2\}] \quad \textit{for any} \quad \gamma > 0. \end{aligned} \tag{4.19}$$

4.8. Proof. We have from (4.14), (4.16), (4.18) that

$$\bar{N}(S, \epsilon) \geqq [\exp\{C - \tfrac{1}{2}\sigma^2\tau - \tfrac{1}{2}(\tau^{-1} - 1)\gamma^2\}],$$

for each choice of $0 < \tau \leqq \frac{1}{2}$. If $\gamma \leqq \frac{1}{2}\sigma$ one can choose $\tau = \gamma/\sigma$, yielding the first part of (4.19). Choosing $\tau = \frac{1}{2}$, one obtains the second part.

Lemma 4.5. *Let b and d be positive constants satisfying $d < b - 1$. Then there exists a positive constant γ_0, depending only on d and b, such that*

$$\bar{N}(S, \epsilon) \leqq \exp\{C - \tfrac{1}{2}\gamma(2\sigma - \gamma) + (\gamma^{-1}b\sigma - d)\log\gamma\} \tag{4.20}$$

holds for each $\gamma \geqq \gamma_0$ and each channel S of normal type (C, σ).

Remark 4.1. Observe that $\gamma^{-1}b\sigma - d < 0$ as soon as $\gamma > b\,d^{-1}\sigma$ where $b\,d^{-1}$ can be arbitrarily close to 1. It follows that the first inequality (4.19) would be false for $\gamma \geqq (1 + \delta)\sigma$ and $\gamma \geqq \gamma_0(\delta)$, no matter how small $\delta > 0$.

4.9. Proof of Lemma 4.5. Let us apply (4.12) with $\theta_2 = C - \sigma z$, where $z = \gamma - \gamma^{-1}b\log\gamma$. Then one obtains

$$\bar{N}(S, \epsilon) \leqq e^{C - \sigma z}[\Phi(z) - \epsilon]^{-1} = [\Phi(z) - \epsilon]^{-1}\exp\{C - \sigma\gamma + \gamma^{-1}b\sigma\log\gamma\}.$$

This yields (4.20) provided $\Phi(z) - \epsilon \geqq \exp\,(-\frac{1}{2}\gamma^2 + d\log\gamma) = \epsilon\gamma^d$. Indeed, we have for all sufficiently large values γ that

$$\begin{aligned} \Phi(-z) > Kz^{-1}e^{-\frac{1}{2}z^2} &\geqq K\gamma^{-1}\exp\{-\tfrac{1}{2}\gamma^2 + b\log\gamma + o(1)\} \\ &= K\,\epsilon\,\gamma^{b-1}e^{o(1)} \geqq \epsilon(1 + \gamma^d), \end{aligned}$$

as soon as γ is sufficiently large, (since $d < b - 1$). Here, K denotes any positive constant with $K < (2\pi)^{-\frac{1}{2}}$.

4.10. The following result is concerned with the case that ϵ is fixed (say $\epsilon = .01$) while σ and hence C are large. Further, $z = \psi(\epsilon)$ is defined by $\Phi(z) = \epsilon$. Thus $\psi(\epsilon)$ is *positive or negative according to whether* $\epsilon > \frac{1}{2}$ *or* $\epsilon < \frac{1}{2}$, while $\psi(\frac{1}{2}) = 0$. Analogous results were obtained by Strassen [23] and Kemperman [11] for the case of a general discrete or semi–continuous stationary memoryless channel.

LEMMA 4.6. *Let* $k > 2$ *be a given constant. Then there exist positive constants* A *and* B *such that*

$$|\log N(S, \epsilon) - C - \psi(\epsilon)\sigma| \leqq \log (A + B\sigma), \tag{4.21}$$

for each $k^{-1} \leqq \epsilon \leqq 1 - k^{-1}$ *and each channel* S *of normal type* (C, σ), (C *and* σ *arbitrary*). *Analogous results hold for* $\bar{N}(S, \epsilon)$.

4.11. *Proof.* If $\sigma = 0$ then F is concentrated at C and (4.12) implies that

$$k^{-1}e^{C} \leqq \epsilon e^{C} \leqq N(S, \epsilon) \leqq \bar{N}(S, \epsilon) \leqq (1 - \epsilon)^{-1}e^{C} \leqq ke^{C}.$$

Thus assume $\sigma > 0$. Let $I = \{z\colon (2k)^{-1} \leqq \Phi(z) \leqq 1 - (2k)^{-1}\}$ and let $a > 0$ be a lower bound on $\Phi'(z)$ when $z \varepsilon I$. Let further $\psi(\epsilon) = \xi_0$ thus $\Phi(\xi_0) = \epsilon$.

Now apply (4.12) with $\theta_i = C + \xi_0\sigma \pm 1 = C + \xi_i\sigma$ $(i = 1, 2)$, where $\xi_1 = \xi_0 - \sigma^{-1}$ and $\xi_2 = \xi_0 + \sigma^{-1}$. One finds that

$$q_1/e \leqq N(S, \epsilon) \exp (-C - \psi(\epsilon)\sigma) \leqq \bar{N}(S, \epsilon) \exp (-C - \psi(\epsilon)\sigma) \leqq e/q_2,$$

where $q_1 = \epsilon - \Phi(\xi_1) = \Phi(\xi_0) - \Phi(\xi_1)$ and $q_2 = \Phi(\xi_2) - \epsilon$.

It suffices to show that $q_i \geqq (A + B\sigma)^{-1}$ for A and B as suitable constants, under the sole assumption that $k^{-1} \leqq \epsilon \leqq 1 - k^{-1}$. As to q_1, we are ready if $\Phi(\xi_1) < (2k)^{-1}$ (in which case $q_1 > (2k)^{-1}$, thus, one may assume that $\xi_1 \varepsilon I$. But then we have from the mean value theorem that $q_1 \geqq a(\xi_0 - \xi_1) = a/\sigma$. Similarly, $q_2 > (2k)^{-1}$ when $\Phi(\xi_2) \geqq 1 - (2k)^{-1}$, while otherwise $q_2 \geqq a/\sigma$.

4.12. *Channels without memory and with additive noise.* Let again S denote a fixed channel with additive noise as described in **4.4**, and let us study the corresponding stationary memoryless channel S^n defined in **2.2**. Clearly, S^n itself is also a channel with additive noise, namely, relative to the n-fold direct product $X^{(n)} = G \times \cdots \times G$ of the compact group $G = X$ corresponding to S. As is easily seen, the distribution function $F^{(n)}(\theta)$ corresponding to S^n is precisely the n-fold convolution $F* \cdots *F$ of the distribution function $F(\theta)$ corresponding to S. Recall that F can be any distribution function satisfying (4.8).

4.13. In the following Lemma Φ_* denotes a distribution function such that F belongs to the domain of partial attraction of Φ_*. This means that one can find an increasing sequence $\{n_k\}$ of positive integers, an increasing sequence $\{\alpha_k\}$ of positive constants with $\alpha_k \to \infty$, and a sequence $\{\beta_k\}$ of real constants, such that

$$\lim_{k\to\infty} F^{(n_k)}(\alpha_k\xi + \beta_k) = \Phi_*(\xi), \tag{4.22}$$

for every continuity point ξ of the function Φ_*.

Let ϵ be given, $0 < \epsilon < 1$. The ϵ-quantile $\psi_*(\epsilon)$ of Φ_* is defined as any number ξ

with $\Phi_*(\xi - 0) \leqq \epsilon \leqq \Phi_*(\xi + 0)$. The set of such values ξ is an interval $[\xi^-, \xi^+]$, where $\xi^- = \sup\{\theta : \Phi_*(\theta) < \epsilon\}$ and $\xi^+ = \inf\{\theta : \Phi_*(\theta) > \epsilon\}$. Consequently, the ϵ-quantile $\xi_0 = \psi(\epsilon)$ is unique if and only if

$$\Phi_*(\xi_1) < \epsilon < \Phi_*(\xi_2) \quad \text{whenever} \quad \xi_1 < \xi_0 < \xi_2. \tag{4.23}$$

Lemma 4.7. *With the above notations, we have*

$$\log N(S^{n_k}, \epsilon) = \beta_k + \alpha_k \psi_*(\epsilon) + o(\alpha_k) \quad \text{as} \quad k \to \infty, \tag{4.24}$$

provided ϵ is such that the ϵ-quantile $\psi_*(\epsilon)$ is unique. A similar result holds for $\bar{N}(S^{n_k}, \epsilon)$.

4.14. Proof. Apply Lemma 4.2 with S replaced by S^{n_k} and F replaced by $F^{(n_k)}$. Choose further $\theta_i = \alpha_k \xi_i + \beta_k$ $(i = 1, 2)$ with $\xi_1 < \xi_0 < \xi_2$ as continuity points of Φ_* arbitrarily close to $\xi_0 = \Phi_*(\epsilon)$. Using (4.22), one immediately obtains (4.24).

4.15. Observe that one can always attain (4.22) with Φ_* concentrated at a single point, namely, by choosing the α_k sufficiently large; thus this so-called trivial domain of partial attraction contains all distribution functions F.

On the other hand, it is quite possible that a particular distribution function belongs to no non-trivial domain of partial attraction. For example, from a result in Feller [8] page 556, this happens when $L(\theta) = 1 - F(\theta)$ is slowly varying as $\theta \to +\infty$, such as $F(\theta) = 1 - \alpha (\log \theta)^{-p}$ for θ large, $(p > 0)$. Note that one can still attain (4.8).

A function $L(\theta)$ is said to be slowly varying as $\theta \to \infty$ when $L(\theta) \neq 0$ for all large θ and $L(k\theta)/L(\theta) \to 1$ as $\theta \to \infty$, for each choice of the constant $k > 0$.

It further follows from a construction in [8] page 557 that there exist functions F which satisfy (4.8) and simultaneously belong to the domain of partial attraction of every infinitely divisible distribution supported by $[0, +\infty)$; (recall that Φ_* in (4.22) is necessarily infinitely divisible).

The above remarks amount to the conclusion that "things can be rather wild," due to the behavior of $F(\theta)$ for large θ. Observe that $F(\theta)$ is necessarily exponentially small as $\theta \to -\infty$, since (4.8) implies that

$$F(-z) \leqq e^{-z} \int_{-\infty}^{-z} e^{-\theta}\, dF(\theta) \leqq e^{-z}, \tag{4.25}$$

4.16. Let us finally consider the more well-behaved channels S, such that the corresponding distribution function F satisfies

$$\lim_{n\to\infty} F^{(n)}(a_n \xi + b_n) = \Phi_*(\xi), \tag{4.26}$$

for all numbers ξ. Here, Φ_* is assumed to be a distribution function not concentrated at a single point. Further, $a_n > 0$ and b_n are suitable constants with $a_n \to +\infty$. It is known [8] that Φ_* is necessarily continuous, in fact, Φ_* is a so-called stable distribution.

4.17. Two distribution functions G and H are said to be of the same type if one can find constants $a > 0$ and b such that $G(a\theta + b) = H(\theta)$ for all θ. In (4.26) the type of Φ_* is uniquely determined by F. All the possible types of Φ_* can be described by a 1-parameter family of distribution functions $\{\Phi_\alpha\,;\, 0 < \alpha \leqq 2\}$ to be described below. It contains the standardized normal distribution function Φ as the special member $\Phi_2 = \Phi$.

Without loss of generality, one may assume that in (4.26) the constants a_n and b_n have been chosen in such a manner that (4.26) holds with $\Phi_* = \Phi_\alpha$, for a unique $0 < \alpha \leqq 2$. Thus, we are interested in channels S such that, for a fixed value α and all ξ,

$$\lim_{n\to\infty} F^{(n)}(a_n\xi + b_n) = \Phi_\alpha(\xi). \tag{4.27}$$

If so we shall say that the channel S is of type Φ_α. Observe that, for $\sigma > 0$, the channels of normal type (C, σ) described in (4.22) are very special channels of type Φ_2. A channel may not be of any of the types Φ_α as follows from the remarks in **4.15**.

Consider a channel of type Φ_α; thus, (4.27) holds for a suitable choice of the a_n and b_n, $(a_n \to +\infty)$. It follows from Lemma 4.7 that

$$\log N(S^n, \epsilon) = b_n + a_n\psi_\alpha(\epsilon) + o(a_n), \qquad \text{as } n \to +\infty, \tag{4.28}$$

holds for each fixed $0 < \epsilon < 1$. Here, $\xi = \psi_\alpha(\epsilon)$ is defined as the unique number ξ with $\Phi_\alpha(\xi) = \epsilon$.

Let

$$C = \textstyle\int_{-\infty}^{+\infty} \theta\, dF(\theta) = \int \{\log f(y)\} f(y)\, dy. \tag{4.29}$$

Then $C \geqq 0$ is finite when $1 < \alpha \leqq 2$, infinite when $0 < \alpha < 1$. It turns out that in (4.28) one can take $b_n = nC$ when $1 < \alpha \leqq 2$, $b_n = 0$ when $0 < \alpha < 1$, (the choice being more complicated when $\alpha = 1$). Finally, it is known that in (4.28) the sequence $\{a_n\}$ must "behave about" as $\{n^{\alpha^{-1}}\}$. Comparing (4.28) with **4.2**, it is clear that C is precisely the Shannon capacity of S; however, there exist much easier proofs of this fact.

4.18. Most of the unproved assertions in the previous section will be rather obvious to a reader familiar with the material in [8], especially Chapters 9 and 17. For completeness, let us now summarize some of the known results concerning (4.26) and (4.27). It will follow from this discussion that *each of the following formulae can be realized* by choosing the additive noise in an appropriate manner.

$$\log N(S^n, \epsilon) - nC \sim cn^{\frac{1}{2}}\psi_2(\epsilon); \tag{4.30}$$

$$\log N(S^n, \epsilon) - nC \sim n^{\frac{1}{2}}(\log n)^3\psi_2(\epsilon); \tag{4.31}$$

$$\log N(S^n, \epsilon) - nC \sim n^{\frac{2}{3}}(\log\log n)^{-12}\psi_{\frac{3}{2}}(\epsilon); \tag{4.32}$$

$$\log N(S^n, \epsilon) \sim n^4\psi_{\frac{1}{4}}(\epsilon). \tag{4.33}$$

Here, the type (4.30) is most common in applications. It holds if and only if F has a finite second moment, hence, certainly when f is bounded, hence, when the group G is finite, hence, when S^n is a so-called binary symmetric channel; namely, take G as a group with two elements. For the latter channel, a result of the type (4.30) was already given by Weiss [24].

The type (4.31) holds, for instance, when $F'(\theta) = k\theta^{-3}(\log \theta)^5$ for θ large, with k as a suitable positive constant. Similarly, (4.32) holds when $F'(\theta) = k\theta^{-\frac{5}{2}}(\log \log \theta)^{-18}$ for θ large, and (4.33) when $F'(\theta) = k\theta^{-\frac{5}{4}}$ for θ large.

4.19. Because of the restriction (4.25) on F not every stable type can arise in (4.26) but only the following 1-parameter family (up to a scale transformation). If $0 < \alpha \leqq 2$ and $\alpha \neq 1$, we define Φ_α by its Fourier transform

$$\int_{-\infty}^{+\infty} e^{is\theta}\, d\Phi_\alpha(\theta) = \exp\{-K_\alpha |s|^\alpha e^{\mp\frac{1}{2}\pi\alpha i}\}. \tag{4.34}$$

Further,

$$\int_{-\infty}^{+\infty} e^{is\theta}\, d\Phi_1(\theta) = \exp\{-K_1 |s| (1 \pm i \log |s|\}. \tag{4.35}$$

Here, s is real while (in $\mp$ and $\pm$) we take the upper sign when $s > 0$, the lower sign when $s < 0$. Further, K_α denotes a fixed constant, chosen once and for all according to convention; some authors take $K_\alpha = 1$. If we want $\Phi_2 = \Phi$ then necessarily $K_2 = \frac{1}{2}$.

If $0 < \alpha < 2$ then $\lim_{\theta \to +\infty} \theta^\alpha[1 - \Phi_\alpha(\theta)] = c_\alpha$, with c_α as a known positive constant. If $0 < \alpha < 1$ then the support of Φ_α is the *half line* $(0, +\infty)$. If $1 \leqq \alpha \leqq 2$ then the support of Φ_α is the *full line* $(-\infty, +\infty)$ but such that $\theta^\alpha \Phi_\alpha(\theta) \to 0$ as $\theta \to -\infty$. It follows that

$$\psi_\alpha(\epsilon) \sim [c_\alpha/(1 - \epsilon)]^{1/\alpha} \quad \text{as} \quad \epsilon \uparrow 1,$$

$(0 < \alpha < 2)$. As $\epsilon \downarrow 0$ we have $\psi_\alpha(\epsilon) \downarrow 0$ when $0 < \alpha < 1$. For $1 \leqq \alpha < 2$ we have $\psi_\alpha(\epsilon) \to -\infty$ in such a way that $\psi_\alpha(\epsilon) = o(\epsilon^{\alpha^{-1}})$.

It is known that (4.27) can occur with $\alpha = 2$ if and only if the function $L_2(\theta) = \int_0^\theta s^2\, dF(s)$ is slowly varying as $\theta \to +\infty$. Further, given $0 < \alpha < 2$, the relation (4.27) is possible if and only if $L_\alpha(\theta) = \theta^\alpha(1 - F(\theta))$ is slowly varying as $\theta \to +\infty$; (recall that we always insist on condition (4.8)).

In both cases, (4.27) can be achieved by means of a sequence $\{a_n\}$ satisfying $a_n \sim k[nL_\alpha(a_n)]^{\alpha^{-1}}$, where k denotes a positive constant depending on F. The b_n may be chosen in the way indicated in the paragraph following (4.29).

The above information gives us a reasonably good grasp of $N(S^n, \epsilon)$ for a large class of channels with additive noise. It would not be difficult to obtain results of the type (4.19) or (4.20). The main advantage of such results would be that they can serve as a source of hints, examples and counterexamples in our explorations of the general continuous channel.

5. Proofs and remarks.

5.1. In proving assertion (2.5), let $\{\epsilon_k\}$ be a fixed sequence of positive constants such that $\epsilon_k \uparrow 1$. Let

$$\{(x_k^{(i)}, D_k^{(i)}); i = 1, \cdots, N(S, \epsilon_k) < \infty\}$$

be an ϵ_k-code for S of *maximal* (but finite) length; $k = 1, 2, \cdots$. Next, let $c_{ki} > 0$, $\sum_{k,i} c_{ki} = 1$, and consider the probability measure

$$\nu(B) = \sum_{k=1}^{\infty} \sum_{i=1}^{N(S,\epsilon_k)} c_{ki} P(B \mid x_k^{(i)}), \qquad (B \,\varepsilon\, \mathfrak{F}_Y).$$

We claim that $P_x \ll \nu$ for each $x \,\varepsilon\, X$.

On the contrary, let $B \,\varepsilon\, \mathfrak{F}_Y$ and $x \,\varepsilon\, X$ be fixed and such that $\nu(B) = 0$ and $P(B \mid x) > 0$. Choose k so large that $P(B \mid x) \geqq 1 - \epsilon_k$.

Since $\nu(B) = 0$ and $c_{ki} > 0$, we have $P(B \mid x_k^{(i)}) = 0$ for all i. Deleting B from each set $D_k^{(i)}$, one may as well assume that the above ϵ_k-code is such that B is disjoint from the $D_k^{(i)}$. But then this code could be enlarged by adding the pair (x, B) and this contradicts the maximality of $N(S, \epsilon_k)$.

5.2. *Proof of Lemma* 3.1. If $N \leqq \bar{N}_L(S, \epsilon)$ then there exists a code of length N as in (2.3) and with $x^{(i)} \,\varepsilon\, L$. Let ν be such that $P_{x_i} \ll \nu$ for all i, and put $f_i = dP_{x_i}/d\nu$, $f_* = \max_i f_i$. The D_i being disjoint, we have

$$\int f_* \, d\nu \geqq \sum_i \int_{D^{(i)}} f_* \, d\nu \geqq \sum_i \int_{D^{(i)}} f_i \, d\nu = \sum_i P(D^{(i)} \mid x^{(i)}) \geqq N(1 - \epsilon).$$

Conversely, let $x^{(i)} \,\varepsilon\, L$ $(i = 1, \cdots, N)$ be such that (3.1) holds, and let f_i and f_* be as above. Then one can attain (2.3) by a suitable choice of the disjoint measurable sets $D^{(i)}$. For instance, one could take $D^{(i)}$ as the set of all $y \,\varepsilon\, Y$ such that $f_* = f_i > \max_{j<i} f_j$.

5.3. *Proof of Lemma* 3.2. Consider a code of length N as in (2.3) having an average error $\leqq \epsilon$. Let

$$E_i = \{y : dP(\cdot \mid x^{(i)})/d\nu \leqq e^{\theta}\}, \qquad F_i = D^{(i)} \cap E_i .$$

Here, the F_i are disjoint, thus,

$$e^{\theta} \geqq \sum_i e^{\theta} \nu(F_i) \geqq \sum_i P(F_i \mid x^{(i)}) \geqq \sum_i \{P(D^{(i)} \mid x^{(i)}) + P(E_i \mid x^{(i)}) - 1\}$$
$$\geqq N(1 - \epsilon) + N(\epsilon + e^{-\rho}) - N = Ne^{-\rho}.$$

Consequently, $N \leqq e^{\theta+\rho}$.

5.4. *Proof of the inequality* (3.5). By homogeneity, we may assume that $\sum f_i = 1$. Suppose $\max(f_1, \cdots, f_N) = f_1$ and put $a_1 = N$, $a_i = 1$ otherwise. Then, from the concavity of $\log z$,

$$f_1 \log N - \sum_{i=1}^{N} \phi(f_i) = \sum_{i=1}^{N} f_i \log(a_i/f_i) \leqq \log \sum_{i=1}^{N} f_i(a_i/f_i)$$
$$\leqq \log \sum_{i=1}^{N} a_i < \log 2N = -N\phi(N^{-1}) + \log 2.$$

Dividing by N, one obtains (3.5).

5.5. *Proof of Lemma* 3.3. It suffices to prove (3.15). We are only interested in points x belonging to the support of Π, (for which $P_x \ll \nu^{\Pi}$). Put

$$f(x, y) = dP(\cdot \mid x)/d\nu^{\Pi}.$$

Let N denote the largest integer for which there exists an ϵ-code as in (2.2) and such that $x^{(i)} \,\varepsilon\, L$ and $D^{(i)} \subset E(x^{(i)})$, $(i = 1, \cdots, N)$. Here,

$$E(x) = \{y : f(x, y) > e^{\theta}\}.$$

We may assume $N < \infty$ (otherwise we are ready). Clearly, $\nu(D^{(i)}) < e^{-\theta}$ thus $\nu(D) < Ne^{-\theta}$, where D stands for the union of the sets $D^{(i)}$.

By the maximality of N, we have for each $x \varepsilon L$ that $P(E(x) \cap D^c | x) < 1 - \epsilon$. But $P(E(x)| x) = 1 - F(\theta | x, \nu^{\Pi})$, thus,

$$P(D | x) > \epsilon - F(\theta | x, \nu^{\Pi}) \quad \text{for each} \quad x \varepsilon L.$$

Integrating over L with respect to Π, and using (3.7) and $Ne^{-\theta} > \nu(D)$, one obtains (3.15).

5.6. *Proof of Lemma* 3.5. Put $a = \sum_{i=1}^{N} E\{Z_i^t\}$. One may assume that $a < \infty$ (otherwise (3.25) would be obvious). Introduce further $p = (1 - t)/t$, thus, $t^{-1} = 1 + p$ and $0 < p \leqq 1$.

The function z^p ($z \geqq 0$) is concave, hence,

$$z^p \leqq a^p + pa^{p-1}(z - a) = rz + s, \qquad (z \geqq 0),$$

where $r = pa^{p-1}$ and $s = (1 - p)a^p$ denote non-negative constants with $ra + s = a^p$.

Further introduce $U_i = \max\{Z_j : j \neq i\}$ and $V_i = \sum_{j \neq i} Z_j^t$. Then

$$W \leqq \sum_{i=1, Z_i \leqq U_i}^{N} Z_i \leqq \sum_{i=1}^{N} Z_i^t U_i^{1-t} \leqq \sum_{i=1}^{N} Z_i^t V_i^p \leqq \sum_{i=1}^{N} Z_i^t (rV_i + s).$$

Here, we used that $V_i^p \geqq \max_{j \neq i} Z_j^{tp} = U_i^{1-t}$. Moreover, by (3.23),

$$E\{Z_i^t V_i\} = \sum_{j \neq i} E\{Z_i^t Z_j^t\} \leqq E\{Z_i^t\} \sum_{j \neq i} E\{Z_j^t\} \leqq E\{Z_i^t\} a.$$

We conclude that

$$E\{W\} \leqq \sum_{i=1}^{N} E\{Z_i^t\}(ra + s) = a(ra + s) = a^{1+p} = a^{t^{-1}},$$

proving (3.25).

Observe that condition (3.23) cannot be omitted entirely. For instance, if $N = 2$ and $Z_1 = Z_2 = Z$ then (3.25) would say that $E\{Z\} \leqq [2E\{Z^t\}]^{t^{-1}}$ for $\frac{1}{2} \leqq t < 1$. Clearly, the latter is false in general.

It is essential that $\frac{1}{2} \leqq t \leqq 1$. For, taking $Z_i = 1$ (all i), (3.25) would say that $N - 1 \leqq N^{t^{-1}}$ which is false for $t > 2$. Taking the Z_i independent,

$$\Pr(Z_i = 1) = \delta, \qquad \Pr(Z_i = 0) = 1 - \delta,$$

(3.25) would imply

$$(N\delta)^{t^{-1}} \geqq E\{W\} \geqq \binom{N}{2}(1 - \delta)^{N-2}\delta^2,$$

which is false when $t < \frac{1}{2}$ and δ is sufficiently small.

5.7. *Proof of an assertion in* **4.4.** Let η and ν be two regular probability measures on the compact group G such that $P_x(B) = \eta(-x + B)$ is absolutely continuous with respect to ν, for each $x \varepsilon X$. We must prove that η is absolutely continuous with respect to the Haar measure μ on G.

Let $D \varepsilon \mathcal{B}$ satisfy $\eta(D) > 0$; we must prove that $\mu(D) > 0$. Observing that $P_x(x + D) = \eta(D) > 0$ it follows from $P_x \ll \nu$ that $\nu(x + D) > 0$ for all $x \varepsilon G$. This in turn implies that $\mu(D) = \int \nu(x + D)\, dx > 0$; (observe that $\lambda(B) =$

$\int \nu(x + B)\, dx$ is a left invariant measure with $\lambda(G) = 1$, thus, λ must coincide with μ).

5.8. *Proof of Lemma* 4.1. For $v > 0$, let $h_v(\theta) = 0$ or $e^{-\theta}$ according to whether $\theta \leqq \log v$ or $\theta > \log v$, respectively. It suffices to prove that for *each* such function h_v we have

$$\int_0^1 h_v(\log f(s))f(s)\, ds = \int_{-\infty}^{+\infty} h_v(\theta)\, dF(\theta).$$

By (4.9), this is equivalent to

$$\int_{0, f(s) > v}^{1} ds = g(v).$$

From their definitions, the function g is continuous to the right, both f and g are non-increasing, while for each $0 < s < 1$ we have $f(s) > v$ if and only if $s < g(v)$. Consequently, the latter integral is equal to min $(1, g(v))$, thus, we need that $g(v) \leqq 1$ for all v. But this is indeed true by (4.8).

6. Entropy. The present section is a slight digression and concerns itself with the precise relations between entropy and total variation.

6.1. Let $Y = (Y, \mathfrak{F}_Y)$ be a fixed measurable space and let μ_1, μ_2 be two probability measures on Y. The entropy $H(\mu_1 \mid \mu_2)$ of μ_1 relative to μ_2 is defined as

$$H(\mu_1 \mid \mu_2) = \sup \textstyle\sum_i \mu_1(B_i) \log \mu_1(B_i)/\mu_2(B_i). \tag{6.1}$$

Here, the supremum is taken over all partitions of Y into finitely many measurable sets B_i ; only the terms with $\mu_1(B_i) > 0$ can make any contribution to the sum, possibly $+\infty$ (namely, when $\mu_2(B_i) = 0$).

Equivalently, (see [19] pages 20 and 24) one has $H(\mu_1 \mid \mu_2) = +\infty$ when μ_1 is not absolutely continuous with respect to μ_2, while otherwise

$$H(\mu_1 \mid \mu_2) = \int \{\log d\mu_1/d\mu_2\}\, d\mu_1 = \int \phi(d\mu_1/d\mu_2)\, d\mu_2\,. \tag{6.2}$$

It follows from the strict convexity of ϕ that we always have $H(\mu_1 \mid \mu_2) \geqq 0$ in such a way that

$$H(\mu_1 \mid \mu_2) > 0 \quad \text{if and only if} \quad \mu_1 \neq \mu_2\,. \tag{6.3}$$

It is convenient to employ a probability measure λ on Y such that both $\mu_1 \ll \lambda$ and $\mu_2 \ll \lambda$, (such as $\lambda = \frac{1}{2}(\mu_1 + \mu_2)$). Put $f_i = d\mu_i/d\lambda$. Then

$$H_{12} = H(\mu_1 \mid \mu_2) = \int \{\log f_1/f_2\} f_1\, d\lambda, \tag{6.4}$$

while

$$H_{21} = H(\mu_2 \mid \mu_1) = -\int \{\log f_1/f_2\} f_2\, d\lambda. \tag{6.5}$$

These representations are always valid. If $\mu_1 \ll \mu_2$ fails to hold then both sides of (6.4) are equal to $+\infty$. Similarly for (6.5).

As is well known (compare (6.3)), both H_{12} and H_{21} may be used as a measure of the distance between μ_1 and μ_2. This is also true for the symmetric distance

$$J_{12} = H_{12} + H_{21} = \int \{\log f_1/f_2\} (f_1 - f_2)\, d\lambda, \tag{6.6}$$

which was already employed by Jeffreys [10] (see [16]). We shall be interested in comparing these distances with the total variation

$$(6.7) \qquad \|\mu_1 - \mu_2\| = \int |f_1(y) - f_2(y)| \lambda(dy)$$

of the signed measure $\mu_1 - \mu_2$. Obviously, $0 \leqq \|\mu_1 - \mu_2\| \leqq 2$. Here, the first equality sign holds if and only if $\mu_1 = \mu_2$, the second equality sign if and only if $\mu_1 \perp \mu_2$.

6.2. It is easily seen that there does not exist any relation between the two numbers H_{12} and H_{21} in the sense that the possible pairs (H_{12}, H_{21}) *fill the entire open positive quadrant* (to which the origin must be added to account for the case $\mu_1 = \mu_2$). Moreover, given the numbers H_{12}, H_{21} the total variation $\|\mu_1 - \mu_2\|$ can be arbitrarily small.

Thus the only remaining problem is to establish upperbounds on $\|\mu_1 - \mu_2\|$ in terms of H_{12} and H_{21}. In this direction, Pinsker ([19] pages 15 and 20) has proved already that there exists an absolute constant $\gamma > 0$ such that

$$(6.8) \qquad \|\mu_1 - \mu_2\| \leqq (\gamma H_{12})^{\frac{1}{2}}$$

holds for every pair of probability measures μ_1, μ_2; (naturally, such an assertion is only of interest when H_{12} is small). Csiszar ([3] page 187) has shown that $\gamma = 16$ will do while McKean ([17] page 358) has proved (6.8) with $\gamma = 4e$.

THEOREM 6.1. *We have*

$$(6.9) \qquad \|\mu_1 - \mu_2\| \leqq (2H_{12})^{\frac{1}{2}}$$

whatever the probability measures μ_1, μ_2. Moreover, the constant $\gamma = 2$ is the best possible.

6.3. *Proof.* To prove the last assertion, let $(Y, \mathfrak{F}_Y, \lambda)$ be any probability space such that λ is not concentrated at one point. Then there exists a bounded measurable function h on this space with $\int h\, d\lambda = 0$, $\int |h|\, d\lambda \neq 0$; one can even attain that $h(y) = \pm 1$.

Choose $\mu_1 = \lambda$, thus, $f_1 = 1$ and let μ_2 be such that $f_2 = 1 - \delta h$, where $\delta \neq 0$ is small. In this case

$$H_{12} = \int \log (1 - \delta h)^{-1}\, d\lambda = \tfrac{1}{2} \int (\delta h)^2\, d\lambda + O(\delta^3).$$

Moreover,

$$\|\mu_1 - \mu_2\| = \int |\delta h|\, d\lambda \leqq \left(\int (\delta h)^2\, d\lambda\right)^{\frac{1}{2}} \sim (2H_{12}).^{\frac{1}{2}}$$

Here, the inequality sign reduces to an equality when $h(y) = \pm 1$ everywhere.

In proving (6.9), one may assume that $\mu_1 \ll \mu_2$, (for, otherwise, $H_{12} = +\infty$). Thus we can take $\lambda = \mu_2$ so that $f_2 = 1$. Observe that $f = f_1 = d\mu_1/d\lambda$ satisfies $\int f\, d\lambda = 1$.

Introducing the *non-negative* (convex) function $\psi(z) = z \log z - z + 1$ on $[0, \infty)$, we see that (6.9) is equivalent to

$$\left[\int |f - 1|\, d\lambda\right]^2 \leqq 2 \int f \log f\, d\lambda = \left[\int \left(\tfrac{4}{3} + \tfrac{2}{3} f\right) d\lambda\right]\left[\int \psi(f)\, d\lambda\right].$$

This in turn would follow from Schwarz's inequality, provided one can show that

$$(f - 1)^2 \leqq (\tfrac{4}{3} + \tfrac{2}{3}f)\psi(f) \tag{6.10}$$

holds for all numbers $f \geqq 0$. In fact, let $\alpha(f)$ denote the difference between the right and left hand sides of (6.10). Then $\alpha(1) = 0$, $\alpha'(1) = 0$, $\alpha''(f) = \tfrac{4}{3}\psi(f)/f \geqq 0$, (a prime denoting differentiation). This proves (6.10) and hence (6.9). The proof also shows that the equality sign in (6.9) cannot hold unless $\mu_1 = \mu_2$.

THEOREM 6.2. *Let $J > 0$ be a given number and let $\rho = \rho(J)$ denote the unique number with $0 < \rho < 1$ and*

$$J = 2\rho \log (1 + \rho)(1 - \rho)^{-1} = 4 \textstyle\sum_{n=1}^{\infty} \rho^{2n}/(2n - 1). \tag{6.11}$$

Then for any pair of probability measures μ_1, μ_2 with $J_{12} = J$ we have

$$\|\mu_1 - \mu_2\| \leqq 2\rho. \tag{6.12}$$

Moreover, for each fixed J, the upperbound (6.12) *is the best possible.*

REMARK 6.1. We have $J > 4\rho^2$ thus $\rho < \tfrac{1}{2}J^{\frac{1}{2}}$ thus

$$\|\mu_1 - \mu_2\| \leqq J_{12}^{\frac{1}{2}} = (H_{12} + H_{21})^{\frac{1}{2}}.$$

Actually, by (6.9), we even have that

$$\|\mu_1 - \mu_2\| \leqq \min\left((2H_{12})^{\frac{1}{2}}, (2H_{21})^{\frac{1}{2}}\right) \leqq (H_{12} + H_{21})^{\frac{1}{2}}.$$

6.4. *Proof of Theorem* 6.2. Let μ_1, μ_2 be such that $J_{12} = J < \infty$. Taking $\lambda = \mu_2$ in (6.6), we have

$$\int (f - 1) \log f \, d\lambda = J, \tag{6.13}$$

where $f = d\mu_1/d\mu_2$.

Suppose that σ and $\tau > 0$ are constants such that

$$\tau |f - 1| \leqq (f - 1) \log f + \sigma(f + 1) \tag{6.14}$$

holds for all numbers $f \geqq 0$. Then (6.13) would imply that

$$\|\mu_1 - \mu_2\| = \int |f - 1| \, d\lambda \leqq (J + 2\sigma)/\tau.$$

Therefore, in proving (6.12), it suffices to establish (6.14) for the special constants

$$\sigma = 2\rho^2(1 - \rho^2)^{-1}; \qquad \tau = (J + 2\sigma)/(2\rho).$$

Here, $0 < \rho < 1$ will be chosen as in (6.11); thus,

$$\tau = \log\{(1 + \rho)(1 - \rho)^{-1}\} + \rho^{-1}\sigma.$$

Since (6.14) has an obvious symmetry (on replacing f by $1/f$), we need to prove only that the quantity $\alpha(f) = (f - 1) \log f + (\sigma - \tau)f + \sigma + \tau$ is non-negative for all $f \geqq 1$. In fact, as $\alpha''(f) = f^{-1} + f^{-2} > 0$ for all $f > 0$, it suffices to show that there exists a (necessarily unique) number $c > 0$ with $\alpha(c) =$

$\alpha'(c) = 0$. It is easily verified that $c = (1 + \rho)(1 - \rho)^{-1}$ will do. This completes the proof of (6.12).

The proof also shows that the equality sign in (6.12) holds if and only if f takes only the values c and c^{-1}. This proves the last assertion. More precisely, to attain the upperbound in (6.12) we can choose μ_1, μ_2 with the same 2-point support (a, b) and such that $\mu_1(a) = \mu_2(b) = \frac{1}{2}(1 + \rho)$; thus, $\mu_1(b) = \mu_2(a) = \frac{1}{2}(1 - \rho)$. Then $J_{12} = 2\rho \log (1 + \rho)(1 - \rho)^{-1}$, while $\|\mu_1 - \mu_2\| = 2\rho$, showing that (6.12) cannot be improved.

Acknowledgment. My sincere thanks to Professor Jack Wolfowitz for introducing me to this area.

REFERENCES

[1] Augustin, U. (1966). Gedächtnisfreie Kanäle für diskrete Zeit. *Z. Wahrscheinlichkeitstheorie* 6 10–61.

[2] Blackwell, D., Breiman, L. and Thomasian, A. J. (1959). The capacity of a class of channels. *Ann. Math. Statist.* **30** 1229–1241.

[3] Csiszar, I. (1966). A note on Jensen's inequality. *Studia Sci. Math. Hungar.* **1** 185–188.

[4] Fano, R. M. (1952). Statistical theory of comunication. Lecture Notes. Massachusetts Institute of Technology, Cambridge.

[5] Fano, R. M. (1961). *Transmission of Information.* M. I. T. Press and Wiley, New York.

[6] Feinstein, A. (1954). A new basic theorem of information theory. *IRE Trans. PGIT* **1** 2–22.

[7] Feinstein, A. (1958). *Foundations of Information Theory.* McGraw-Hill, New York.

[8] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

[9] Gallager, R. G. (1965). A simple derivation of the coding theorem and some applications. *IEEE Trans. Information Theory* **IT-11**, 3–18.

[10] Jeffreys, H. (1948). *Theory of Probability,* 2nd ed. Oxford University Press, Oxford.

[11] Kemperman, J. H. B. (1960). Upper and lower bounds on the length of the longest code (abstract). *Notices Amer. Math. Soc.* **7** 924.

[12] Kemperman, J. H. B. (1962). *Studies in coding theory* I. Mimeograph Report, Univ. of Rochester.

[13] Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. *Proc. Internat. Symposium Probability and Information Theory,* M. Behara, ed. Springer, New York.

[14] Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory.* Dover, New York.

[15] Kotz, S. (1966). *Recent Results in Information Theory.* Methuen, London.

[16] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22** 79–86.

[17] McKean Jr., H. P. (1966). Speed of approach to equilibrium for Kac's caricature of a Maxwellian gas. *Arch. Rational Mech. Anal.* **21** 343–367.

[18] McMillan, B. (1953). The basic theorems of information theory. *Ann. Math. Statist.* **24** 196–219.

[19] Pinsker, M. S. (1964). *Information and Information Stability of Random Variables and Processes,* A. Feinstein, tr. and ed. Holden-Day, San Francisco.

[20] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423; 623–656.

[21] SHANNON, C. E. (1957). Certain results in coding theory for noisy channels. *Information and Control* **1** 6–25.

[22] SHANNON, C. E., GALLAGER, R. G. and BERLEKAMP, E. R. (1967). Lower bounds to error probability for coding on discrete memoryless channels I. *Information and Control* **10** 65–103.

[23] STRASSEN, V. (1964). Asymptotische Abschätzungen in Shannons Informationstheorie. *Trans. Third Prague Conference on Information Theory.* Publ. House Czechoslovak Acad. Sc., Prague. 1–35.

[24] WEISS, L. (1960). On the strong converse of the coding theorem for symmetric channels without memory. *Quart. Appl. Math.* **18** 209–214.

[25] WOLFOWITZ, J. (1957). The coding of messages subject to chance errors. *Illinois J. Math.* **1** 591–606.

[26] WOLFOWITZ, J. (1959). Strong converse of the coding theorem for semicontinuous channels. *Illinois J. Math.* **3** 477–489.

[27] WOLFOWITZ, J. (1961). *Coding Theorems of Information Theory.* Springer, New York.

[28] WOLFOWITZ, J. (1964). *Coding Theorems of Information Theory*, 2nd ed. Springer, New York.