SIAM REVIEW
Vol. 11, No. 3, July 1969

ON CODING AND INFORMATION THEORY*

A. D. WYNER†

1. Introduction. In this introductory paper we give a survey of some important coding theorems of information theory and some related problems. Particular emphasis has been placed on the so-called "packing problem" for the binary n-cube where there are several important open questions. No effort was made to make this survey a comprehensive one, but rather several important areas are treated in considerable detail.

To fix ideas let us consider the following communication problem: A data source emits a sequence of statistically independent equally likely binary digits (0's and 1's) at a rate R per second. These digits are to be communicated over an unreliable channel which transmits binary digits correctly with probability $1 - p_0$, and incorrectly with probability p_0. In other words, if the channel input is a 0 (or 1) the channel output is 0 (or 1) with probability $1 - p_0$, and 1 (or 0) with probability p_0. The channel is *memoryless* in the sense that each input digit is perturbed independently. This channel is known as the *binary symmetric channel.* We shall assume that the channel passes one digit per second.

If we set the data rate $R = 1$, then the source can be connected directly to the channel and the digits will be communicated with error probability $P_e = p_0$. We may improve the reliability of the system by setting $R = 1/n$ and repeating each transmission n times, where $n = 2m + 1$ is an odd integer. If the number of received 0's exceeds the number of received 1's in a block of n received digits, the receiver announces a 0 and vice versa. Thus an error occurs when $m + 1$ or more transmissions out of n are incorrect. This event has probability

$$P_e = \sum_{k \geq m+1} \binom{n}{k} p_0^k (1 - p_0)^{n-k}. \tag{1.1}$$

If $p_0 < \frac{1}{2}$, the error probability $P_e \to 0$ as $n \to \infty$. The penalty which we must pay for this increased reliability is that the rate $R = 1/n$ also tends to zero.

We now pose the question: Is it possible to communicate over this channel at a fixed positive rate R with arbitrarily small error probability? The answer was given by C. E. Shannon in his now classic paper *A Mathematical Theory of Communication* published in 1948 [1], and is the essence of information theory. Roughly speaking, Shannon's answer to our question is "yes," provided R does not exceed a quantity called the "channel capacity." The means by which this is done is a process called "coding," in which the channel input digits are made to depend not on a single data digit (as in our example), but on many data digits in a complex way. Arbitrarily small error probability P_e is attainable in the limit as the coding complexity becomes large. We outline Shannon's result (called the "fundamental theorem of information theory") below.

* Received by the editors March 28, 1968.

† Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey 07974. On leave, 1969–1970, at Weizmann Institute of Science, Rehovot, Israel.

FIG. 1

Let us modify the communication system to that shown in Fig. 1. The *source*, exactly as before, emits a sequence of independent equally likely binary digits at a rate of R per second. The *channel*, as before, is the binary symmetric channel with parameter p_0, which passes one digit per second. The *coder*, which is interposed between the source and the channel, waits n seconds during which time a sequence of $[Rn]$[1] digits will have collected at the coder input. We call such a sequence of digits a "message." Thus there are $M = 2^{[Rn]}$ possible messages, each message equally likely. Let us number them $1, 2, \cdots, M$. The coder contains a mapping of the message set $\{1, 2, \cdots, M\}$ to a set of M binary n-sequences $\{\mathbf{x}_i\}, i = 1, \cdots, M$ (called a *code*). If message i, $i = 1, 2, \cdots, M$, is the coder input, then the coder output is $\mathbf{x}_i$. Since the channel takes n seconds to pass $\mathbf{x}_i$, the system can process data continuously without a "backup" at the coder input. In the above example where each source digit is repeated n times, the code consists of the two n-sequences $(0, 0, \cdots, 0)$ and $(1, 1, \cdots, 1)$.

It is the task of the *decoder* to examine the channel output $\mathbf{y}$, which is a possibly corrupted version of $\mathbf{x}_i$, and to decide which message was actually transmitted. Thus the decoder is a mapping D of the space of binary n-sequences back to the message set $\{1, 2, \cdots, M\}$. An "error" is said to occur whenever $D(y) \neq i$, where i is the transmitted message. Let us denote by P_{ei}, the error probability given that message i is transmitted, $i = 1, 2, \cdots, M$. The overall error probability is then

$$P_e = \frac{1}{M} \sum_{i=1}^{M} P_{ei}. \tag{1.2}$$

We can now state Shannon's "fundamental theorem:"

For $0 < \rho \leqq \frac{1}{2}$, define the quantity[2]

$$C_B(\rho) = 1 + \rho \log_2 \rho + (1 - \rho) \log_2 (1 - \rho). \tag{1.3}$$

Then for any $R < C_B(p_0)$, and any $\varepsilon > 0$, there exist an n (sufficiently large) and a coding-decoding scheme with parameters n and R for the binary symmetric channel with parameter p_0 such that $P_e \leqq \varepsilon$. Further if $R \geqq C_B(p_0)$, there exists an $\varepsilon = \varepsilon(R) > 0$ such that for any coding-decoding scheme with parameters n and R, $P_e \geqq \varepsilon$ no matter how large n may be.

In other words, $C_B(p_0)$ is the supremum of those source rates R for which it is possible to communicate over the binary symmetric channel with arbitrarily high reliability as the code dimension n approaches infinity. Consequently $C_B(p_0)$ is called the *channel capacity*. (A graph of $C_B(p_0)$ is given in Fig. 2. Note that $C_B(p_0)$ decreases from 1 to 0 as p_0 increases from 0 to $\frac{1}{2}$. Thus when $p_0 = 0$, the channel is "noiseless" and the capacity is unity, and when $p_0 = \frac{1}{2}$, the channel is "useless" and has zero capacity.) The price which must be paid for the high reliability which can be attained through coding are (i) a fixed delay which occurs

[1] The symbol $[x]$ is the largest integer not exceeding x.

[2] We also define $C_B(0) = \lim_{\rho \to 0} C_B(\rho) = 1$.

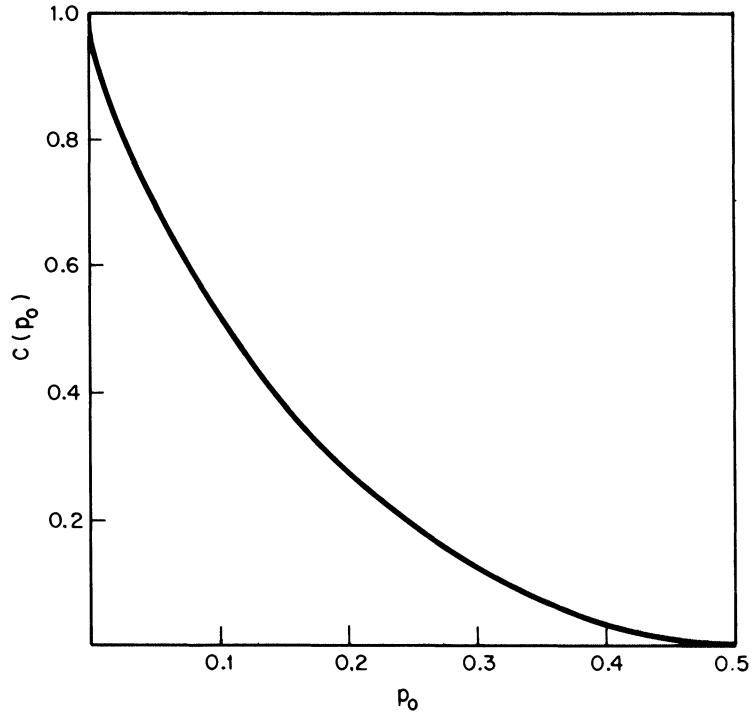

FIG. 2. $C(p_0)$ vs p_0

because the coder and decoder must each initially wait n seconds to begin their respective functions and (ii) coding-decoding equipment complexity. Roughly speaking these costs increase with increasing n.

Let us comment at this point that Shannon's remarkable results do not completely solve the communication problem. In fact Shannon opened a treasure house of problems, many of the most enticing of which have remained unsolved. These open problems center about the fact that Shannon only established the *existence* of coding-decoding schemes with certain desirable properties, but he did not show explicitly how to construct the coder and decoder. In fact, despite enormous effort, an explicit method of generating the coder-decoder promised by Shannon still eludes information theorists to this day.

In § 2 we shall discuss in detail the coding problem for the binary symmetric channel as well as some closely related problems. In the course of this discussion we shall give a proof of the fundamental theorem. Special emphasis will be put on the geometric aspects of the problem, particularly on the so-called "packing problem" on the binary n-cube. In § 3 we shall consider the coding problem for another channel, the so-called "Gaussian channel" where the input is a real number and the output is the sum of the input and a Gaussian "noise" variate. The Gaussian channel, like the binary symmetric channel, also lends itself to geometric analysis. In § 3 we shall state without proof several results which parallel those for the binary symmetric channel. Finally, in § 4 we describe a very general class of channels and give a coding theorem about them.

Before beginning the main part of the paper let us take a moment to mention some general references on our subject. We begin with Shannon's classic paper [1]. This readable paper founded information theory and contains most of the important ideas in the subject. Probably the most comprehensive work on information theory is a new book by R. G. Gallager [2].[3] It is complete, mathematically correct, and well written, and is essential reading for any serious student of this subject. The subject of coding theorems is covered fully by Wolfowitz [3]. Less complete but perhaps more easily read treatments can be found in Ash [4] and Abramson [5]. Finally we mention the monograph by Kotz [6], which is unusually strong on the contribution by Soviets to this subject, and which contains an extensive bibliography.

2. The binary symmetric channel.

2.1. Statement of the principal problem and the fundamental theorem. Shannon's fundamental theorem tells us only what is possible asymptotically as the block length n approaches infinity for a fixed transmission rate R. A more complete statement of the coding problem is as follows. Let $P_e^*(M, n, p_0)$ be the smallest attainable error probability (defined by (1.2)) for an n-dimensional code with M code words used on a binary symmetric channel with parameter p_0. A code which achieves $P_e^*(M, n, p_0)$ is called *optimal*. We now state the coding problem.

PROBLEM 2.1. For arbitrary M, n and p_0, what is $P_e^*(M, n, p_0)$, and what is (are) the code(s) which achieves it?

Surprisingly, comparatively little is known about the quantity $P_e^*(M, n, p_0)$, and still less about the codes which achieve it. A trivial case is that of $M = 2$. Here the code given in the introductory example in § 1 is optimal for all n and p_0, and $P_e^*(2, n, p_0)$ is given by (1.1) for odd n and by a similar expression for even n. We shall give an example of nontrivial optimal codes presently (in § 2.4).

Shannon's fundamental theorem is an asymptotic result about $P_e^*(M, n, p_0)$.

THEOREM 2.1 (Shannon). *Let $M = M_n = 2^{[Rn]}$, where the rate R is held fixed as $n \to \infty$. Then*

(2.1) (i)[4] *if $R < C_B(p_0)$, $P_e^*(M_n, n, p_0) \xrightarrow{n} 0$;*

(ii) *if $R \geqq C_B(p_0)$, there exists a $\lambda = \lambda(R, p_0) > 0$ such that $P_e^*(M_n, n, p_0) \geqq \lambda$ for all n,*

where

$$C_B(\rho) = 1 + \rho \log_2 \rho + (1 - \rho) \log_2 (1 - \rho), \qquad 0 \leqq \rho \leqq \tfrac{1}{2}. \tag{2.2}$$

Part (i) of the theorem, called the "direct half," is proved below. We postpone the proof of part (ii), called the "converse half," to § 2.3.

Proof of the direct half. Our strategy is as follows:

---

[3] After the preparation of this manuscript another comprehensive book *Probabilistic Information Theory* by F. Jelinek (McGraw-Hill) was released.

[4] A sharper result is also true: Let $R < C_B(p_0)$ be held fixed and let $M_n = 2^{[Rn]}$. Then $P_e^*(M_n, n, p_0) = \exp\{-E(R, p_0)n(1 + \varepsilon_1(n))\}$, where $\varepsilon_1 \to 0$ as $n \to \infty$, and $E(R, p_0) > 0$. Thus at rates R less than capacity, P_e approaches zero exponentially in n. Although the exponent $E(R, p_0)$ is not known exactly for all R and p_0, rather tight upper and lower bounds for this quantity are available. See [7], [8].

(a) Let $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, be a code with parameters M and n for use on the binary symmetric channel with parameter p_0. We shall find a function $Q(\mathbf{x}_1, \cdots, \mathbf{x}_M)$ of the M code words such that the error probability attainable with this code is overbounded by Q. Certainly $P_e^*(M, n, p_0) \leqq Q(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M)$ for all codes.

(b) Consider an experiment in which M (binary) random n-vectors $\mathbf{X}_1, \cdots, \mathbf{X}_M$ are chosen independently with identical distribution

$$\Pr\{\mathbf{X}_k = \mathbf{x}\} = 2^{-n} \quad \text{for all } \mathbf{x}. \tag{2.3}$$

Then $Q(\mathbf{X}_1, \cdots, \mathbf{X}_M)$ (where Q is defined in (a)) is a random variable. Let $\bar{Q}$ be its expectation. $\bar{Q}$ is a function of M, n and p_0. Certainly $P_e^*(M, n, p_0) \leqq \bar{Q}(M, n, p_0)$. We then show that $\bar{Q}(2^{[Rn]}, n, p_0) \to 0$ as $n \to \infty$ when $R < C_B(p_0)$ which establishes the direct half.

We now proceed with the proof.

(a) Let $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, be a code with parameters M and n for use on the binary symmetric channel with parameter p_0. Consider the following decoding rule.

Let $\mathbf{u}$ be a binary n-vector, and for all α, $0 \leqq \alpha \leqq 1$, define the "sphere" $S_\alpha(\mathbf{u}) = \{\mathbf{v}: d_H(\mathbf{u}, \mathbf{v}) \leqq \alpha n\}$, where $d_H(\mathbf{u}, \mathbf{v})$ (the "Hamming distance" between the binary n-vectors $\mathbf{u}$ and $\mathbf{v}$) is the number of coordinate positions in which $\mathbf{u}$ and $\mathbf{v}$ differ. The decoding rule D is as follows: Let ρ, $0 < \rho < 1$, be a number to be specified later. If $\mathbf{y}$ is the received vector and *exactly one* code word, say $\mathbf{x}_j$, belongs to $S_\rho(\mathbf{y})$, then $D(\mathbf{y}) = j$. Otherwise the decoder makes no attempt to decode $\mathbf{y}$ and declares an error. Note that since with very high probability roughly $p_0 n$ errors will occur in transmission, one would expect that it would be wise to choose ρ greater than p_0. Further the error probability P_e which we obtain using this decoding rule cannot be less than the error probability which is obtained for this code with the best decoding rule,[5] which in turn cannot be less than $P_e^*(M, n, p_0)$.

Say $\mathbf{x}_i$, $i = 1, 2, \cdots, M$, is transmitted and $\mathbf{y}$ is received. Then a decoding error will be made only if $\mathbf{x}_i \notin S_\rho(\mathbf{y})$ or $\mathbf{x}_j \in S_\rho(\mathbf{y})$ for some $j \neq i$. Thus the error probability given that code word i is transmitted is

$$P_{ei} = \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}_i)\Phi_i(\mathbf{y}, \{\mathbf{x}_j\}_{j=1}^M), \tag{2.4a}$$

where $p(\mathbf{y}|\mathbf{x}_i)$ is the probability that the received vector is $\mathbf{y}$ given that $\mathbf{x}_i$ is transmitted,

$$\Phi_i(\mathbf{y}, \{\mathbf{x}_j\}_{j=1}^M) = \begin{cases} 1 & \text{if } d_H(\mathbf{x}_i, \mathbf{y}) > \rho n \\ & \text{or } d_H(\mathbf{x}_j, \mathbf{y}) \leqq \rho n \quad \text{for some } j \neq i, \\ 0 & \text{otherwise,} \end{cases} \tag{2.4b}$$

and the summation is taken over all 2^n binary n-vectors $\mathbf{y}$. Defining $\Phi_0(\mathbf{u}, \mathbf{v})$ ($\mathbf{u}, \mathbf{v}$ binary n-vectors) as 0 or 1 according as $d_H(\mathbf{u}, \mathbf{v}) > \rho n$ or $d_H(\mathbf{u}, \mathbf{v}) \leqq \rho n$, we have

$$\Phi_i(\mathbf{y}, \{\mathbf{x}_j\}_{j=1}^M) \leqq [1 - \Phi_0(\mathbf{y}, \mathbf{x}_i)] + \sum_{j \neq i} \Phi_0(\mathbf{y}, \mathbf{x}_j),$$

[5] It is easy to show that the decoding rule which minimizes P_e (for a given code) takes $D(\mathbf{y})$ as that i which minimizes $d_H(\mathbf{y}, \mathbf{x}_i)$.

and (2.4a) becomes

$$\begin{aligned} P_{ei} &\leqq \sum_{\mathbf{y}} [1 - \Phi_0(\mathbf{y}, \mathbf{x}_i)] p(\mathbf{y}|\mathbf{x}_i) + \sum_{\mathbf{y}} \sum_{j \neq i} \Phi_0(\mathbf{y}, \mathbf{x}_j) p(\mathbf{y}|\mathbf{x}_i) \\ &= \sum_{\mathbf{y} \notin S_\rho(\mathbf{x}_i)} p(\mathbf{y}|\mathbf{x}_i) + \sum_{\mathbf{y}} \sum_{j \neq i} \Phi_0(\mathbf{y}, \mathbf{x}_j) p(\mathbf{y}|\mathbf{x}_i). \end{aligned} \tag{2.5}$$

Consider the first summation of (2.5). This is merely the probability that the received vector $\mathbf{y}$ and the transmitted vector $\mathbf{x}_i$ differ in more than ρn positions. Since the digits are corrupted independently by the channel, this quantity is

$$\sum_{\rho n < r \leqq n} \binom{n}{r} p_0^r (1 - p_0)^{n-r} \triangleq \psi_n(\rho, p_0). \tag{2.6}$$

Also since $P_e = (1/M) \sum_{i=1}^{M} P_{ei}$, we have from (2.5) and (2.6) the desired bound

$$P_e^*(M, n, p_0) \leqq Q_\rho(\mathbf{x}_1, \cdots, \mathbf{x}_M), \tag{2.7a}$$

where

$$Q_\rho(\mathbf{x}_1, \cdots, \mathbf{x}_M) = \psi_n(\rho, p_0) + \frac{1}{M} \sum_{i=1}^{M} \sum_{\mathbf{y}} \sum_{j \neq i} p(\mathbf{y}|\mathbf{x}_i) \Phi_0(\mathbf{y}, \mathbf{x}_j). \tag{2.7b}$$

(b) We now imagine that the n-vectors $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, are chosen at random independently with identical distributions given by (2.3), and find the expectation $\bar{Q}_\rho(M, n, p_0)$ of Q_ρ. Letting E denote expectation we have from (2.7b)

$$\begin{aligned} \bar{Q}_\rho(M, n, p_0) &= E Q_\rho(\mathbf{X}_1, \cdots, \mathbf{X}_M) \\ &= \psi_n(\rho, p_0) + \frac{1}{M} \sum_{i=1}^{M} \sum_{j \neq i} \sum_{\mathbf{y}} E\{p(\mathbf{y}|\mathbf{X}_i) \Phi_0(\mathbf{y}, \mathbf{X}_j)\}, \end{aligned} \tag{2.8}$$

where upper-case $\mathbf{X}$'s are used to represent random vectors. Now since for $i \neq j$ the random vectors $\mathbf{X}_i$ and $\mathbf{X}_j$ are independent, for any fixed $\mathbf{y}$ the random variables $p(\mathbf{y}|\mathbf{X}_i)$ and $\Phi_0(\mathbf{y}, \mathbf{X}_j)$ are independent, and therefore

$$E\{p(\mathbf{y}|\mathbf{X}_i) \Phi_0(\mathbf{y}, \mathbf{X}_j)\} = E p(\mathbf{y}|\mathbf{X}_i) \cdot E \Phi_0(\mathbf{y}, \mathbf{X}_j).$$

Now with $\mathbf{y}$ fixed, $\Phi_0(\mathbf{y}, \mathbf{X}_j)$ is a random variable which assumes the value 1 when $\mathbf{X}_j \in S_\rho(\mathbf{y})$ and 0 otherwise. Thus

$$\begin{aligned} E\Phi_0(\mathbf{y}, \mathbf{X}_j) &= \Pr\{\mathbf{X}_j \in S_\rho(\mathbf{y})\} = \frac{\text{Card } S_\rho(\mathbf{y})}{2^n} \\ &= \sum_{0 \leqq r \leqq \rho n} \binom{n}{r} 2^{-n} \triangleq \xi_n(\rho). \end{aligned} \tag{2.9}$$

Thus the second term of (2.8) is

$$\frac{1}{M} \sum_{i=1}^{M} \sum_{j \neq i} \xi_n(\rho) E\left(\sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{X}_i) \right).$$

Since the random variable $\sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{X}_i)$ is identically unity, so is its expectation, and we can write

$$\bar{Q}_\rho(M, n, p_0) = \psi_n(\rho, p_0) + (M - 1)\xi_n(\rho), \tag{2.10}$$

where ψ_n is defined in (2.6) and ξ_n in (2.9).

Now set $M = 2^{[Rn]}$, where $R < C_B(p_0)$. Since C_B is a continuous monotone decreasing function, we can write $R = C_B(p_0 + \lambda_R)$, $\lambda_R > 0$. We now show that if we set $\rho = p_0 + (\lambda_R/2)$ then $\bar{Q}_\rho(2^{[Rn]}, n, p_0) \to 0$ as $n \to \infty$, which completes the proof. Consider the first term of (2.10). This is the probability of more than $n\rho$ "successes" in a sequence of Bernoulli trials with "success" probability p_0. By the weak law of large numbers [9], $\psi_n(\rho, p_0) \xrightarrow{n} 0$ when $\rho = p_0 + (\lambda_R/2)$. Now consider the second term of (2.10). It follows directly from the Stirling formula [9, pp. 50–52] or [4, p. 113, Eq. 4.7.1] that for $0 \leqq \rho \leqq \frac{1}{2}$,

$$\xi_n(\rho) = \sum_{0 \leqq r \leqq \rho n} \binom{n}{r} 2^{-n} = 2^{-nC_B(\rho) + O(\log n)}, \tag{2.11}$$

where $C_B(\rho)$ is defined in (2.2). Thus with $\rho = p_0 + (\lambda_R/2)$,

$$(M - 1)\xi_n(\rho) \leqq 2^{-n[C_B(p_0 + \lambda_R/2) - C_B(p_0 + \lambda_R)] + o(n)} \to 0 \quad \text{as } n \to \infty. \tag{2.12}$$

This completes the proof of the direct half of the fundamental theorem.

2.2. Hamming distance and the sphere-packing bound. An important concept related to the coding problem is the so-called "minimum distance principle." Let $\mathbf{u}$ and $\mathbf{v}$ be binary n-sequences, and as above let the (Hamming) distance $d_H(\mathbf{u}, \mathbf{v})$ be the number of positions in which $\mathbf{u}$ and $\mathbf{v}$ differ. It is easily verified that d_H is in fact a metric. Further let $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, be a code and let $d = \min_{i \neq j} d_H(\mathbf{x}_i, \mathbf{x}_j)$. Then it follows from the metric property of d_H (triangle inequality[6]) that the code is capable of correcting all errors which corrupt $(d - 1)/2$ or fewer coordinates of the transmitted vector. An intuitive way of thinking about this is to consider the M (closed) spheres of Hamming radius $(d - 1)/2$ about each code vector. These spheres are disjoint. Then, if the channel corrupts no more than $(d - 1)/2$ coordinates of the transmitted vector $\mathbf{x}_i$, the received vector will lie in the sphere about code word $\mathbf{x}_i$ and the error can be corrected. We might also suppose that if, for a given M and n, it were possible to subdivide the binary n-cube into M disjoint spheres of equal radius that contain all the 2^n points of the cube, then the code consisting of the centers would be optimal. This is in fact the case, and we state this formally as a theorem which is a useful lower bound to $P_e^*(M, n, p_0)$. In fact, this bound leads us directly to the converse half of the fundamental theorem.

Let n be a positive integer and let α, $1 \leqq \alpha \leqq 2^n$, be a real number. Let $r_0 = r_0(\alpha, n)$ be the largest integer such that

$$\sum_{k=0}^{r_0} \binom{n}{k} \leqq \alpha. \tag{2.13}$$

[6] The triangle inequality is $d_H(\mathbf{u}, \mathbf{v}) \leqq d_H(\mathbf{u}, \mathbf{w}) + d_H(\mathbf{w}, \mathbf{v})$.

Then let $G(\alpha) = G(\alpha, n, p_0)$ be

$$G(\alpha) = \sum_{k=0}^{r_0} \binom{n}{k} p_0^k (1-p_0)^{n-k} + \left(\alpha - \sum_{k=0}^{r_0} \binom{n}{k} \right) p_0^{r_0+1} (1-p_0)^{n-r_0-1}. \tag{2.14}$$

THEOREM 2.2 (Sphere-packing bound). *For arbitrary* $n, p_0, M \leqq 2^n$,

$$P_e^*(M, n, p_0) \geqq 1 - G(2^n/M). \tag{2.15}$$

A case of special interest is that when $M = 2^n/(n+1)$, so that $r_0((2^n/M), n) = 1$. Inequality (2.15) is then $P_e^*(M, n, p_0) \geqq 1 - (1-p_0)^n - np_0(1-p_0)^{n-1}$. The optimal class of codes which we promised to give later will satisfy this bound with equality.

Proof of Theorem 2.2. Let $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, be an optimal code with parameters M and n, and let D be the decoder mapping which achieves error probability $P_e^*(M, n, p_0)$. Let N_i, $i = 1, 2, \cdots, M$, be the number of vectors $\mathbf{y}$ for which $D(\mathbf{y}) = i$. The probability of error given that code word i is transmitted, P_{ei}, is at least the probability of error if these N_i vectors were placed as closely as possible to code word $\mathbf{x}_i$ (since $p(\mathbf{y}|\mathbf{x}_i)$, the probability of receiving vector $\mathbf{y}$ given that $\mathbf{x}_i$ is transmitted, is a decreasing function of $d_H(\mathbf{y}, \mathbf{x}_i)$). The closest they could be placed to $\mathbf{x}_i$ is in a sphere of radius $r_0(N_i, n)$ with the remainder on a "shell" of radius $r_0 + 1$. If this were possible, $P_{ei} = 1 - G(N_i)$. Thus in general

$$P_{ei} \geqq 1 - G(N_i). \tag{2.16}$$

It is easily shown that $G(\alpha)$ is a concave (downward) function of α so that

$$\begin{aligned} P_e^*(M, n, p_0) &= \frac{1}{M} \sum_{i=1}^{M} P_{ei} \\ &\geqq 1 - \sum_{i=1}^{M} \frac{1}{M} G(N_i) \\ &\geqq 1 - G\left(\frac{\sum N_i}{M}\right) = 1 - G(2^n/M), \end{aligned} \tag{2.17}$$

as required.

We now apply Theorem 2.2.

2.3. Proof of the converse half of the fundamental theorem. Let $R \geqq C_B(p_0)$. We show that, for a binary symmetric channel with parameter p_0, $P_e^*(2^{[Rn]}, n, p_0)$ is bounded away from zero. In fact, we shall show that

$$P_e^*(2^{[Rn]}, n, p_0) \xrightarrow{n} 1, \qquad R > C_B(p_0) \tag{2.18}$$

and $P_e^*(2^{[C_B(p_0)n]}, n, p_0) \xrightarrow{n} \frac{1}{2}$. Equation (2.18) is called the "strong converse."

Let $M = 2^{[Rn]}$, $R \geqq C_B(p_0)$. Assume that $R \leqq 1$. Since $C_B(\lambda)$ is a continuous decreasing function of λ, $\lambda < \frac{1}{2}$, we can write $R = C_B(\lambda_R)$, $\lambda_R \leqq p_0$. Let $\alpha = 2^n/M$. We have from (2.11) and (2.13) that $r_0(\alpha, n) = \lambda_R n + O(\log n)$. Thus from (2.14), $G(\alpha)$ approaches $\frac{1}{2}$ when $\lambda_R = p_0$ and 0 when $\lambda_R > p_0$. The converse now follows from Theorem 2.2 for $R \leqq 1$. Since $P_e^*(M, n, p_0)$ is nondecreasing in M, the converse is valid for $R > 1$ also.

2.4 Optimal Hamming codes and group codes. We now describe the only known large class of optimal codes (called *Hamming codes* after their discoverer [10]). Let $m > 0$ be an arbitrary integer. The dimension of the codes is $n = 2^m - 1$. Let $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n$ be the n distinct nonzero binary m-vectors. Then the code consists of those binary n-vectors $\mathbf{x} = (x_1, \cdots, x_n)$ which satisfy

$$\sum_{k=1}^{n} x_k \mathbf{u}_k = \mathbf{0} \pmod 2, \tag{2.19}$$

where the addition is ordinary vector addition modulo 2. Since (2.19) represents m linearly independent equations in the n "unknowns" $x_1, \cdots, x_n$, we may choose $k = n - m$ of these coordinates freely (the other m coordinates being constrained by (2.19)) so that the code has $M = 2^k$ code words. Let us observe that the code defined by (2.19) has the property that if $\mathbf{x}_1$ and $\mathbf{x}_2$ are code vectors, then $\mathbf{x}_1 + \mathbf{x}_2$ (mod 2) is also a code vector. We now show that the minimum distance of the Hamming code is 3, so that the code corrects all single errors. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be code vectors and let $d_H(\mathbf{x}_1, \mathbf{x}_2) = d'$. Then $\mathbf{x}_3 = \mathbf{x}_1 + \mathbf{x}_2$ (mod 2) is a code vector with exactly d' nonzero coordinates, say $k_1, k_2, \cdots, k_{d'}$. Thus from (2.19),

$$\sum_{i=1}^{d'} \mathbf{u}_{k_i} = \mathbf{0} \pmod 2.$$

Since the $\mathbf{u}_k$ are distinct and nonzero we conclude that $d' \geqq 3$. Since there are many triples (k_1, k_2, k_3) such that

$$\mathbf{u}_{k_1} + \mathbf{u}_{k_2} + \mathbf{u}_{k_3} = \mathbf{0} \pmod 2,$$

the vector with 1's in the k_1, k_2, k_3 positions (and zeros elsewhere) is a code word which is distance three from the all zero vector (also a code word). Thus the minimum distance is three and the code corrects all single errors. The error probability is therefore

$$\begin{aligned} P_e &\leqq 1 - \Pr\left\{\begin{array}{c}\text{no error in transmission}\\ \text{or}\\ \text{exactly one error in transmission}\end{array}\right\} \\ &= 1 - (1 - p_0)^n - n(1 - p_0)^{n-1} p_0, \end{aligned}$$

which meets the lower bound of Theorem 2.2.

Hamming codes are members of an important class of codes called *parity check* or *group* codes (the latter name following from the fact that the code vectors form a group under modulo 2 vector addition). These codes may be defined by a set of n distinct nonzero m-vectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n$, where $n \leqq 2^m - 1$. The code words are those n-vectors $\mathbf{x}$ which satisfy (2.19). As above, the number of code words is $M = 2^{n-m}$ (provided (2.19) is a set of m linearly independent equations). Using the same reasoning as above, we see that the minimum distance is the size of the minimum subset of the $\mathbf{u}_k$ which sum to zero. It is known that the fundamental theorem (Theorem 2.1) remains valid if we reduce the class of possible codes to group codes. Although no class of group codes is known which has the property guaranteed by this theorem, large and relatively powerful classes of group codes with arbitrary minimum distance are known. For a full discussion of group codes

and other related "algebraic" coding techniques the reader is referred to Peterson [11] or Berlekamp [12]. Less complete but somewhat more easily read discussions may be found in Ash [4] and Gallager [2].

2.5. The packing problem. As is clear from the above, we can overbound the probability of error for a given code by a decreasing function of the minimum distance between code words. It thus seems a reasonable question to ask: for a given M and n what is the largest minimum distance attainable for a code with parameters M and n? Alternatively we pose the problem as the determination of $M^*(n, d)$ defined as the maximum number of code words which may be placed on the binary n-cube with minimum Hamming distance d between code words. A second problem is that of actually finding a code which achieves this maximum.

If we are to use our codes on a binary symmetric channel with parameter p_0, then from the law of large numbers we expect roughly $p_0 n$ errors to occur when n is large. Thus it is quite reasonable to let $d = \beta n$, where β, $0 \leqq \beta \leqq 1$, is a fixed parameter, and consider the behavior of $M^*(n, \beta n)$ as n becomes large. We show below with almost trivial arguments that, with β held fixed, there exist positive constants A and B such that

$$2^{nA} \leqq M^*(n, \beta n) \leqq 2^{nB}.$$

Thus $M(n, \beta n)$ increases exponentially in n. Despite the near triviality of this fact, no explicit construction of a code with minimum distance βn and of exponentially increasing size is known.

In the following two sections (2.5.1 and 2.5.2) we establish upper and lower bounds respectively on $M^*(n, d)$. In § 2.5.3 we use these results to estimate the rate of exponential increase of $M^*(n, \beta n)$. The exact determination of this rate is another of the important open problems in coding theory. In § 2.5.4 we prove an interesting theorem on random packing.

2.5.1. Lower bound on $M^*(n, d)$. Let $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, be a maximum size n-dimensional code with minimum distance d and $M = M^*(n, d)$ code words. Let $\mathscr{S}_i$ be the open sphere in the binary n-cube with radius d and center at $\mathbf{x}_i$, i.e.,

$$\mathscr{S}_i = \{\mathbf{u} : d_H(\mathbf{u}, \mathbf{x}_i) < d\}.$$

If $\mathbf{u}$ (an n-vector) does not belong to $\bigcup_{i=1}^{M} S_i$, then $d_H(\mathbf{u}, \mathbf{x}_i) \geqq d$, $i = 1, \cdots, M$, and $\mathbf{u}$ can be added to the code contradicting its assumed maximality. Thus we conclude that all n-vectors belong to at least one of the spheres $\mathscr{S}_i$ and that

$$\begin{aligned} 2^n = \operatorname{card} \bigcup_{i=1}^{M} \mathscr{S}_i &\leqq \sum_{i=1}^{M} \operatorname{card}(\mathscr{S}_i) \\ &= M \sum_{k=0}^{d-1} \binom{n}{k}, \end{aligned}$$

or

$$M^*(n, d) \geqq 2^n \Big/ \left(\sum_{k=0}^{d-1} \binom{n}{k} \right). \tag{2.20}$$

Inequality (2.20) is due to Gilbert [13] and despite its simplicity is essentially the strongest known general lower bound on $M^*(n, d)$. A slightly sharper version, known as the Gilbert–Varsharmov–Sachs bound, can be found in [11]. This sharper version has the advantage that it asserts the existence of codes which satisfy the lower bound and which are also group codes.

2.5.2. Upper bounds on $M^*(n, d)$. We give here several upper bounds on $M^*(n, d)$. The first, due to Hamming [10], is obtained directly from the observation that given a code with minimum distance d, the set of closed spheres of radius $[(d-1)/2]$ centered at the code words are disjoint. Hence

$$M^*(n, d) \leqq 2^n \Big/ \sum_{k=0}^{[(d-1)/2]} \binom{n}{k}. \tag{2.21}$$

The Hamming codes described above achieve this upper bound for the case $d = 3$, $n = 2^m - 1$, $m = 1, 2, 3 \cdots$. Note that a slight strengthening of (2.21) is possible when d is even

The second upper bound due to Plotkin [14] is given in two parts, valid in the ranges $2d/n > 1$ and $2d/n \leqq 1$ respectively. The first part is

$$M^*(n, d) \leqq 2d/(2d - n), \qquad 2d/n > 1. \tag{2.22}$$

To establish (2.22), let $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, be a code with parameters M, n, d. Write $\mathbf{x}_i = (x_{i1}, \cdots, x_{in})$, where $x_{ik} = 0$ or 1, and consider the array

$$\begin{array}{l} \mathbf{x}_1 = x_{11}, x_{12}, \cdots, x_{1n} \\ \quad\vdots \qquad\quad \vdots \qquad\qquad \vdots \\ \mathbf{x}_M = x_{M1}, \qquad \cdots, x_{Mn} \end{array} \tag{2.23}$$

Denote by s_k the number of 1's in column k of the array. Since the code has minimum distance d, $d_H(\mathbf{x}_i, \mathbf{x}_j) \geqq d$, $i \neq j$, so that

$$\begin{aligned} \binom{M}{2} d &\leqq \sum_{1 \leqq i < j \leqq M} d_H(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{n} s_k(M - s_k) \\ &\leqq \sum_{k=1}^{n} \frac{M^2}{4} = \frac{M^2 n}{4}, \end{aligned} \tag{2.24}$$

where the second inequality follows from the fact that $Mu - u^2 \leqq M^2/4$, $-\infty < u < \infty$, with equality if and only if $u = M/2$. Inequality (2.24) is equivalent to (2.22).

We can achieve equality in (2.22) if the code is such that $d_H(\mathbf{x}_i, \mathbf{x}_j) = d$ for all $i \neq j$, and all the columns of (2.23) are half 0's and half 1's. This is possible when $M = 2k$ and $n = \binom{2k}{k} \cdot j$; $j, k = 1, 2, \cdots$, so that the columns of (2.3) can be j repetitions of the $\binom{2k}{k}$ columns with exactly k 1's (and k 0's). In this case

$$d = \frac{jk}{(2k-1)} \binom{2k}{k},$$

so that

$$M^*(n, d) = 2k, \quad \text{when} \quad n = j\binom{2k}{k}, \tag{2.25}$$

and

$$d = \frac{jk}{(2k-1)}\binom{2k}{k}, \qquad j, k = 1, 2, \cdots.$$

Plotkin also extended this bound to the case $2d/n \leqq 1$ as follows. Consider an n_0-dimensional code with minimum distance d and $M = M^*(n_0, d)$ code words. Partition the code words into two classes according to whether the first coordinate is 0 or 1. One class has at least $M/2$ members, and when the first coordinate of each code word in that class is deleted, the class becomes an $(n_0 - 1)$-dimensional code with minimum distance at least d. Thus $M^*(n_0 - 1, d) \geqq \frac{1}{2}M^*(n_0, d)$ or $M^*(n_0, d) \leqq 2M^*(n_0 - 1, d)$. By induction we obtain for $k = 0, 1, 2, \cdots$

$$M^*(n_0, d) \leqq 2^k M^*(n_0 - k, d). \tag{2.26}$$

Now say we are given n, d, where $n \geqq 2d$, and we want to bound $M(n, d)$ (note that (2.22) does not hold). Set $k_0 = n - 2d + 1$ so that $n - k_0 < 2d$ and apply (2.22) with n replaced by $n - k_0$,

$$M^*(n - k_0, d) \leqq \frac{2d}{2d - (n - k_0)} = 2d. \tag{2.27}$$

Now apply (2.26), with $k = k_0$ and $n_0 = n$, and obtain

$$M^*(n, d) \leqq 2^{k_0}(2d) = d2^{n-2d+2}, \qquad n \geqq 2d. \tag{2.28}$$

Plotkin gave several examples for which the upper bound (2.28) can be realized. In particular, he showed that if $4k - 1$ is a prime, $M^*(4k, 2k) = 8k$, $k = 1, 2, \cdots$.

We now give a final upper bound, valid in the range $2d/n < 1$. This bound is usually, though not always, sharper than Plotkin's bound (2.28). Let n, d satisfy $2d/n < 1$ and let $K = K(n, d)$ be the set of positive integers k for which

$$k < \frac{n}{2}\left[1 - \left(1 - \frac{2d}{n}\right)^{1/2}\right].$$

Then for $2d/n < 1$,

$$M^*(n, d) \leqq \min_{k \in K(n,d)} \frac{2^n}{\binom{n}{k}} \frac{d}{[d - 2k(n-k)/n]}. \tag{2.29}$$

Inequality (2.29) is a variant of bounds found by Elias [15] and Wyner [16]. The form and elegant derivation given here are due to L. A. Shepp. We proceed to the proof of (2.29) beginning with some definitions.

Let $\mathbf{u}$ be a binary n-vector, let the weight of $\mathbf{u}$, denoted $|\mathbf{u}|$, be the number of nonzero entries in $\mathbf{u}$. Let $\hat{M}(n, d, k)$ be the maximum size of an n-dimensional code with minimum distance d such that the weight of all code words is k. Our result follows directly from two lemmas.

LEMMA 1. *For* $k \leqq n$,

$$M^*(n, d) \leqq \frac{2^n}{\binom{n}{k}} \hat{M}(n, d, k), \qquad k = 0, 1, \cdots, n.$$

LEMMA 2. *If* $2d/n < 1$ *and* $k \in K(n, d)$, *then*

$$\hat{M}(n, d, k) \leqq \frac{d}{[d - 2k(n - k)/n]}.$$

Inequality (2.29) follows when k is chosen to yield the tightest bound.

Proof of Lemma 1. Let $k \leqq n$ be a nonnegative integer and let $\mathbf{y}$ be a binary n-vector. Let $\mu_k(\mathbf{y})$ be defined by

$$\mu_k(\mathbf{y}) = \begin{cases} 1 \Big/ \binom{n}{k}, & |\mathbf{y}| = k, \\ 0 & \text{otherwise}. \end{cases} \tag{2.30}$$

Let $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, be an n-dimensional code with M code words and minimum distance d. With k fixed, for each n-vector $\mathbf{y}$ let

$$\mu^*(\mathbf{y}) = \mu_k^*(\mathbf{y}) = \sum_{i=1}^{M} \mu_k(\mathbf{y} + \mathbf{x}_i),$$

where the addition is vector addition modulo 2. Then

$$\begin{aligned} \sum_{\mathbf{y}} \mu^*(\mathbf{y}) = \sum_{i=1}^{M} \sum_{\mathbf{y}} \mu_k(\mathbf{y} + \mathbf{x}_i) &= \sum_{i=1}^{M} \frac{1}{\binom{n}{k}} \left\{ \begin{matrix} \text{no. of } n\text{-vectors} \\ \text{distance } k \text{ from } \mathbf{x}_i \end{matrix} \right\} \\ &= M \end{aligned} \tag{2.31}$$

$\left(\sum_{\mathbf{y}} \text{ is to be taken over all of the } 2^n \text{ } n\text{-vectors}\right)$. Further

$$\sum_{\mathbf{y}} \mu^*(\mathbf{y}) = \sum_{\mathbf{y}} \sum_{i} \mu_k(\mathbf{y} + \mathbf{x}_i) \leqq 2^n \max_{\mathbf{y}} \sum_{i} \mu_k(\mathbf{y} + \mathbf{x}_i). \tag{2.32}$$

Let $\mathbf{y}_0$ achieve the indicated maximum. Then

$$\sum_{i} \mu_k(\mathbf{y}_0 + \mathbf{x}_i) = \frac{1}{\binom{n}{k}} \left\{ \begin{matrix} \text{no. of } \mathbf{x}_i \text{ such that} \\ |\mathbf{x}_i + \mathbf{y}_0| = k \end{matrix} \right\}. \tag{2.33}$$

Since $\{\mathbf{x}_i + \mathbf{y}_0\}$, $i = 1, 2, \cdots, M$, is a code with minimum distance d, the quantity in braces in (2.33) is $\leqq \hat{M}(n, d, k)$. Thus from (2.31), (2.32) and (2.33) we have

$$M \leqq \frac{2^n}{\binom{n}{k}} \hat{M}(n, d, k),$$

which is Lemma 1.

Proof of Lemma 2. Say we are given an n-dimensional code $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, with M code words, minimum distance d, such that $|\mathbf{x}_i| = k$, $i = 1, 2, \cdots, M$. Arrange the code words in an array as in (2.23) and let s_r, $r = 1, 2, \cdots, n$, be the number of 1's in column r of this array. Since $d_H(\mathbf{x}_i, \mathbf{x}_j) \geqq d$, $i \neq j$, we have, as in (2.24),

$$\binom{M}{2} d \leqq \sum_{1 \leqq i < j \leqq M} d_H(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{n} s_r(M - s_r) = M \sum_r s_r - \sum_r s_r^2. \tag{2.34}$$

Now $\sum_r s_r = \sum_{i=1}^{M} |\mathbf{x}_i| = Mk$. Further, from the Schwarz inequality,

$$\left(\sum_{r=1}^{n} s_r \cdot 1\right)^2 \leqq \sum_r s_r^2 \cdot \sum_r 1 = n \sum_r s_r^2,$$

so that

$$\sum s_r^2 \geqq M^2k^2/n.$$

Thus (2.34) yields

$$\binom{M}{2} d \leqq M^2k - M^2k^2/n,$$

or

$$M \leqq \frac{d/n}{d/n - (2k/n)(1 - k/n)},$$

provided $d/n - (2k/n)(1 - k/n) > 0$. The last condition is equivalent to $k \in K(n, d)$. This is Lemma 2.

2.5.3. Asymptotic form of the bounds on $M^*(n, d)$. Let $R^*(n, d) = (1/n) \cdot \log_2 M^*(n, d)$ be the maximum attainable transmission rate for an n-dimensional code with minimum distance d. We are concerned here with the asymptotics of $R^*(n, d)$ for large n. If we are to use our codes on the binary symmetric channel with parameter p_0, then from the law of large numbers we would expect roughly $p_0 n$ errors. Thus it is quite reasonable to let $d = \beta n$, where β, $0 \leqq \beta \leqq 1$, is a fixed parameter. Let us therefore define

$$R^*(\beta) = \limsup_{n \to \infty} R^*(n, \beta n). \tag{2.35}$$

(It is an open question whether $R^*(n, \beta n)$ actually approaches a limit.) Thus, roughly speaking, $R^*(\beta)$ is the largest rate for which it is possible (as $n \to \infty$) to obtain codes with $d/n = \beta$. Using the fact that for fixed ρ, $0 \leqq \rho \leqq 1$,

$$\lim_{n \to \infty} \frac{1}{n} \log_2 \binom{n}{\rho n} = 1 - C_B(\rho) \tag{2.36}$$

(which follows from (2.11)), we obtain from (2.20)

$$R^*(\beta) \geqq C_B(\beta). \tag{2.37}$$

Our upper bounds yield the following: Inequality (2.21) yields

(2.38a) $$R^*(\beta) \leqq C_B(\beta/2),$$

(2.22) yields

(2.38b) $$R^*(\beta) = 0, \qquad \beta > \tfrac{1}{2},$$

(2.28) yields

(2.38c) $$R^*(\beta) \leqq 1 - 2\beta, \qquad \beta \leqq \tfrac{1}{2},$$

and (2.29) yields

(2.38d) $$R^*(\beta) \leqq C_B(\tfrac{1}{2} - \tfrac{1}{2}\sqrt{1 - 2\beta}), \qquad \beta < \tfrac{1}{2}.$$

Since (2.38d) is uniformly sharper than (2.38a) and (2.38c), we have

(2.39a) $$R^*(\beta) = 0, \qquad \beta \geqq \tfrac{1}{2},$$

and

(2.39b) $$C_B(\beta) \leqq R^*(\beta) \leqq C_B(\tfrac{1}{2} - \tfrac{1}{2}\sqrt{1 - 2\beta}), \qquad 0 < \beta < \tfrac{1}{2}.$$

These bounds are plotted in Fig. 3. An exact determination of $R^*(\beta)$ is one of the important outstanding problems of coding theory.

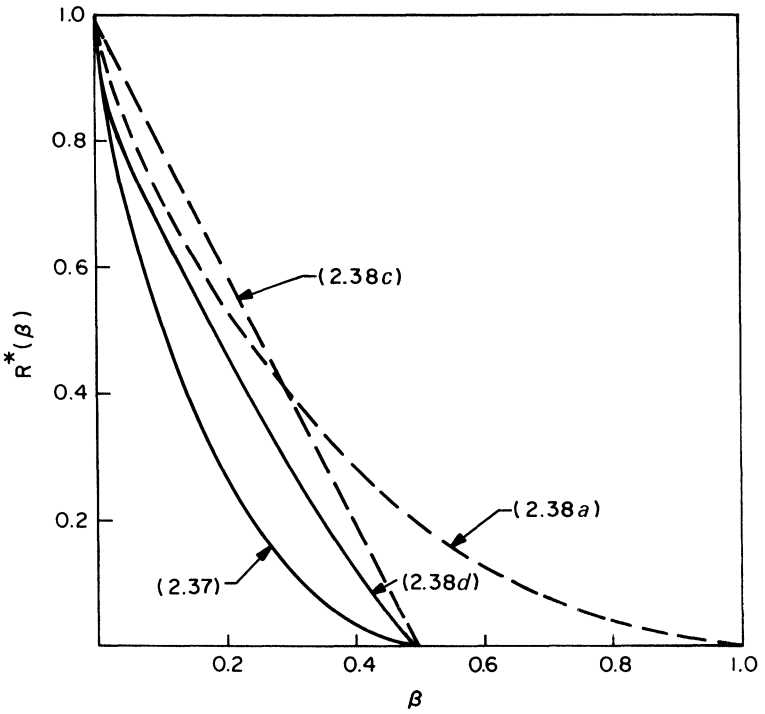

FIG. 3. *Upper and lower bounds on* $R^*(\beta)$ *vs* β

2.5.4. Minimum distance of a random code. We conclude this section with a theorem on the minimum distance of a random code which states roughly that if a code with $2^{nC_B(\beta)}$ members is chosen at random, then on the average (when n is large) only a very small fraction of the code words will have a neighbor closer than βn. We proceed as follows. Let $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, be an n-dimensional code. Let β, $0 \leqq \beta \leqq \frac{1}{2}$, be fixed, and define the functions g_i, $i = 1, \cdots, M$,

$$g_i(\mathbf{x}_1, \cdots, \mathbf{x}_M) = \begin{cases} 1 & \text{if } d_H(\mathbf{x}_i, \mathbf{x}_j) > \beta n \quad \text{for all } j \neq i, \\ 0 & \text{otherwise.} \end{cases} \tag{2.40a}$$

Let

$$F(\mathbf{x}_1, \cdots, \mathbf{x}_M) = \frac{1}{M} \sum_{i=1}^{M} g_i. \tag{2.40b}$$

Clearly the code has minimum distance greater than βn if and only if $F = 1$. We now state the following theorem.

THEOREM 2.3. *Consider a random experiment in which the M members of a code $\{\mathbf{X}_i\}$, $i = 1, 2, \cdots, M$, are chosen independently and uniformly from the set of binary n-vectors. Then $F(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M)$ is a random variable with expectation EF. With β, $0 \leqq \beta \leqq \frac{1}{2}$, held fixed, let M increase as $n \to \infty$. Then if*

$$\frac{M}{2^n} \sum_{k=0}^{\beta n} \binom{n}{k} \to \infty, \quad \textit{then } EF \to 0, \tag{2.41a}$$

and if

$$\frac{M}{2^n} \sum_{k=0}^{\beta n} \binom{n}{k} \to 0, \quad \textit{then } EG \to 1. \tag{2.41b}$$

In particular if $M = 2^{\rho n}$ (ρ fixed) we have (using (2.11)) that $EF \to 0$ or 1 according as $\rho > C_B(\beta)$ or $\rho < C_B(\beta)$. Further, since there must be a code for which $F \geqq EF$ we conclude that for any $\rho < C_B(\beta)$ and any $\varepsilon > 0$, there exist n sufficiently large and an n-dimensional code with $M = 2^{\rho n}$ members such that $F > 1 - \varepsilon$. If we delete the εM members of the code for which $g_i = 0$, we obtain a code with minimum distance βn with $2^{n\rho}(1 - \varepsilon)$ points. This is equivalent to the lower bound of (2.39b), $R^*(\beta) \geqq C_B(\beta)$. To those who have faith in the power of random coding (witness the proof of the fundamental theorem) the present theorem is a plausibility argument for the conjecture that $R^*(\beta)$ is in fact equal to $C_B(\beta)$.

Proof of Theorem 2.3. From (2.40) we have

$$EF = \frac{1}{M} \sum_{i=1}^{M} Eg_i = \frac{1}{M} \sum_{i=1}^{M} \Pr\{g_i = 1\}. \tag{2.42}$$

Let i be fixed. If $\mathbf{X}_i = \mathbf{x}$, then $g_i = 1$ if and only if the $M - 1$ independent choices of $\{\mathbf{X}_j\}$, $j = 1, 2, \cdots, M$, $j \neq i$, satisfy $d_H(\mathbf{X}_j, \mathbf{x}) > \beta n$. Since the $\mathbf{X}_j$ are uniformly distributed on the set of binary n-vectors,

$$\Pr\{g_i = 1 | \mathbf{X}_i = \mathbf{x}\} = \left(1 - \frac{\sum_{k \leqq \beta n} \binom{n}{k}}{2^n}\right)^{M-1}, \tag{2.43}$$

independent of $\mathbf{x}$. Thus from (2.42),

$$EF = \left(1 - \sum_{k<\beta n} \binom{n}{k} 2^{-n}\right)^{M-1} = \left(1 - \frac{1}{u_n}\right)^{u_n((M-1)/u_n)}, \tag{2.44}$$

where $u_n = \left(\sum_{k<\beta n} \binom{n}{k} 2^{-n}\right)^{-1} \to \infty$, as $n \to \infty$.[7] Our result (2.41) follows on noting that as $n \to \infty$, $(1 - 1/u_n)^{u_n} \to e^{-1}$, and $(M-1)/u_n \sim M \sum_{k<\beta n} \binom{n}{k} 2^{-n}$.

3. The Gaussian channel.

3.1. Definitions. In this section we shall treat another channel, called the *Gaussian channel*, defined as follows. The channel input is a real n-vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ which satisfies an "energy" constraint

$$\|\mathbf{x}\|^2 \triangleq \sum_{k=1}^{n} x_k^2 = A. \tag{3.1}$$

Thus $\mathbf{x}$ may be thought of as a point on the surface of a sphere of radius $\sqrt{A}$ in Euclidean n-space. In § 3.6 we shall give an interpretation of $\mathbf{x}$ and $\|\mathbf{x}\|$ in terms of continuous-time signals which can be used to realize $\mathbf{x}$ in a physical system. The channel output is a real n-vector $\mathbf{y} = (y_1, \cdots, y_n)$ which is the vector sum of $\mathbf{x}$ and a noise vector $\mathbf{z} = (z_1, \cdots, z_n)$. The components of $\mathbf{z}$ are independent zero mean Gaussian variates with variance σ^2. As in the case of the binary symmetric channel, a *code* is a set of M n-vectors $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, which satisfy (3.1), and a *decoder* is a mapping D from Euclidean n-space R^n to the set $\{1, 2, \cdots, M\}$. When $\mathbf{x}_i$ is transmitted, an error is made if $\mathbf{y}$ is such that $D(\mathbf{y}) \neq i$. We let P_{ei} be the probability of error when $\mathbf{x}_i$ is transmitted so that

$$P_e = \frac{1}{M} \sum_{i=1}^{M} P_{ei}$$

is the overall error probability when all M code words are equally likely to be transmitted. We denote by $P_e^*(M, n, A, \sigma^2)$ the smallest attainable P_e for a code with parameters M, n and A, and a channel with parameter σ^2.

3.2. The fundamental theorem for the Gaussian channel. Say that the Gaussian channel is to be used in the communication system of Fig. 1. Let us assume that the channel accepts real numbers at a rate of one per second so that it takes n seconds to pass an input n-vector. Further we assume that the allowable signal energy A is proportional to the time it takes to transmit $\mathbf{x}_i$, so that we set $A = Pn$, where P is the "energy per second" or "average power." The ratio $P/\sigma^2 \triangleq \rho$ is called the "signal-to-noise" (power) ratio. Since the source (Fig. 1) emits a sequence of independent equally likely binary digits, the number of possible messages after n seconds is $M = 2^{[Rn]}$. Thus we are concerned with $P_e^*(2^{[Rn]}, n, \rho n, \sigma^2)$ with R, ρ, σ^2 fixed. Shannon's fundamental theorem applies here too, and we state it here without proof.[8] A proof of the direct half is contained in § 4.

[7] Provided $\beta < \frac{1}{2}$. If $\beta = \frac{1}{2}$, (2.41a) always holds and from (2.44), $EF \to 0$, so that the theorem is true in this case also.

[8] Proofs can be found in [17], [18], [4], [3], [2].

THEOREM 3.1 (Shannon). *Suppose we are given a Gaussian channel with parameters P and σ^2. Let $\rho = P/\sigma^2$ and*

$$C_G(\rho) = \tfrac{1}{2}\log_2(1+\rho). \tag{3.2}$$

Then

(i) *if $R < C_G(\rho)$,*

$$P_e^*(2^{[Rn]}, n, Pn, \sigma^2) \to 0 \quad \text{as } n \to \infty; \tag{3.3a}$$

(ii) *if $R \geqq C_G(\rho)$, there exists a $\lambda = \lambda(\rho, R) > 0$ such that*

$$P^*(2^{[Rn]}, n, Pn, \sigma^2) \geqq \lambda, \quad \text{for all } n. \tag{3.3b}$$

Thus, as for the binary symmetric channel, $C_G(\rho)$ is the largest rate at which information can be transmitted with arbitrarily high reliability over the Gaussian channel with signal-to-noise ratio ρ as the delay n becomes infinite. Therefore, C_G is called the *channel capacity.*

3.3. Statement of the principal problem. As for the binary symmetric channel, the principal problem is the determination of $P_e^*(M, n, A, \sigma^2)$ and the discovery of optimal codes. Since it is easy to show that for a given M and n, P_e^* depends on A and σ^2 only through their ratio, we can, without loss of generality, take $A = 1$ and write $P_e^* = P_e^*(M, n, \sigma^2)$, where σ^2 is now the ratio of noise variance to signal energy.

There is one class of codes which has been shown to be optimal [19].[9] We state this in the following theorem.

THEOREM 3.2. *Let $M = n + 1$. Then for all σ^2, $P_e^*(n+1, n, \sigma^2)$ is attained when the code vectors are the $n+1$ vertices of the n-dimensional regular simplex*[10] *inscribed in a unit n-sphere.*

An explicit though somewhat complex expression for $P_e^*(n+1, n, \sigma^2)$ was found by Nuttal [20].

In fact, the following theorem shows that the simplex is optimal whenever $M \leqq n + 1$. The theorem follows directly on observing that any set of M points in n-space must lie in a hyperplane of dimension $M - 1$ or less.

THEOREM 3.3. *Let $\{\mathbf{x}_i\}$, $i = 1, 2, \cdots, M$, be a code with parameters M and n, where $n \geqq M - 1$, with error probability P_e. Then there exists a code $\{\hat{\mathbf{x}}_i\}$, $i = 1, 2, \cdots, M$, with M code words and dimension $\hat{n} \leqq M - 1$ with error probability $\hat{P}_e \leqq P_e$.*

We conclude from Theorems 3.2 and 3.3 that the outstanding problem is the determination of $P_e^*(M, n, \sigma^2)$ where $M > n + 1$. Again, as for the binary symmetric channel, very little is known about P_e^*. Weber [21] has found code configurations which are locally optimal[11] for $n + 2 \leqq M \leqq 2n$. The code for $M = 2n$ is the particularly simple "biorthogonal code" which consists of n orthogonal unit n-vectors and their negatives. In § 3.4 a lower bound on $P_e^*(M, n, \sigma^2)$ is stated which is analogous to the sphere-packing bound for the binary symmetric channel.

[9] After the preparation of this manuscript an error in the proof of Theorem 3.2. in [19] was discovered (S. Farber, Doctoral thesis, Calif. Inst. of Technology, 1968) which to the present author's knowledge has not been resolved. Thus Theorem 3.2 must be considered a conjecture.

[10] In other words, the inner product of all pairs of the $n + 1$ unit code vectors is $-1/n$.

[11] That is any small perturbation of the code points increases P_e.

3.4. Sphere-packing lower bound on P_e^*. Let Ω_0 be the solid angle[12] of a hypersphere in n-space. Let $\theta_k(n)$, $k = 1, 2, \cdots$, be the half-angle of a right circular cone with apex at the center of the hypersphere, which cuts a solid angle Ω_0/k. If it were possible to subdivide n-space into M congruent cones with apex at the origin, placing the M code words on the axis of the cones would result in an error probability $1 - Q(\theta_M(n), n, \sigma^2)$, where $Q(\theta, n, \sigma^2)$ is the probability that the noise (spherical Gaussian with variance σ^2) does not perturb a point $\mathbf{x}$ (on the surface of a unit n-sphere with center at the origin) outside a right circular cone of half-angle θ with apex at the origin and axis the semi-infinite half-line connecting the origin and $\mathbf{x}$. In fact, this quantity is a lower bound on the error probability.

THEOREM 3.3. (Sphere-packing bound).

$$P_e^*(M, n, \sigma^2) \geqq 1 - Q(\theta_M(n), n, \sigma^2).$$

The proof of Theorem 3.3 is nearly identical with that of Theorem 2.2 with angle used instead of Hamming distance.

There is one trivial case when this bound can be realized with equality. This is when $n = 2$, $M = 3, 4, \cdots$. Here we merely take our (optimal) code to be any set of M equally spaced points on the unit circle. For $n \geqq 3$, it is doubtful that such a "perfect" code is achievable.

Theorem 3.3 can be used with the aid of the law of large numbers (as was done for the binary symmetric channel) to establish the converse half of the fundamental theorem (Theorem 3.1(ii)).

3.5. Bounds on attainable packings. As for the binary symmetric channel it is desirable to have codes with a guaranteed large (Euclidean) distance between code words. Since we have assumed that all code words are on the surface of an n-sphere, an equivalent requirement is a large angle between code points. Accordingly, let us define $M^*(n, \theta)$ as the maximum number of points that can be placed on the surface of an n-sphere with minimum angle θ between points (a "packing"). Note that $M^*(n, \theta)$ does not depend on the radius of the sphere.

For $\pi/2 \leqq \theta \leqq 2\pi$, $M^*(n, \theta)$ is known exactly as well as the code configuration which achieves it. This result is due to Rankin [22].

THEOREM 3.4. *For* $\pi/2 \leqq \theta \leqq 2\pi$, $M^*(n, \theta)$ *is*

(i) $M^*(n, \theta) = 1, \quad \pi < \theta \leqq 2\pi,$

(ii) $M^*(n, \theta) = \left[\dfrac{\cos\theta - 1}{\cos\theta}\right]$ *for* $\cos^{-1}(-1/n) \leqq \theta \leqq \pi,$

(iii) $M^*(n, \theta) = n + 1$ *for* $\pi/2 < \theta \leqq \cos^{-1}(-1/n),$

(iv) $M^*(n, \pi/2) = 2n.$

We remark that the code configuration which achieves (ii) and (iii) is a regular simplex, and the configuration which achieves (iv) is the biorthogonal code (see § 3.3). The proof of the upper bound on M^* for (i) and (ii) closely parallels that of Plotkin bound (2.22).

For $0 < \theta < \pi/2$, $M^*(n, \theta)$ is not known exactly, but, as for the binary codes, can be estimated by upper and lower bounds.

[12] The solid angle of a region on the surface of an n-sphere is the ratio of the area of the region to nth power of the radius of the hypersphere.

THEOREM 3.5. *For* $0 < \theta < \pi/2$,

$$\frac{n}{n-1}\sqrt{\pi}\frac{\Gamma((n+1)/2)}{\Gamma((n+2)/2)}\left[\int_0^{\theta}\sin^{n-2}\varphi\,d\varphi\right]^{-1}$$

$$\leqq M^*(n,\theta) \leqq \frac{\sqrt{\pi}}{2}\frac{\Gamma((n-1)/2)}{\Gamma(n/2)}\frac{\sin\psi\sin\psi}{\int_0^{\psi}(\sin\varphi)^{n-2}(\cos\varphi-\cos\psi)\,d\varphi},$$

where $\psi = \sin^{-1}\sqrt{2}\sin(\theta/2)$.

The upper bound was also obtained by Rankin [22] and is analogous to the bound for binary codes of (2.29). The lower bound closely parallels the Gilbert bound of (2.20) and is proved in [16].

Just as for binary codes we let $R^*(n,\theta) = (1/n)\log_2 M^*(n,\theta)$ and $R^*(\theta) = \limsup R(n,\theta)$ as $n \to \infty$. Theorems 3.4 and 3.5 yield estimates of the asymptotic rate $R^*(\theta)$. From Theorem 3.4,

$$R^*(\theta) = 0, \qquad \pi/2 \leqq \theta \leqq 2\pi, \tag{3.4a}$$

and from Theorem 3.5 it is easy to show that

$$-\log_2\sin\theta \leqq R^*(\theta) \leqq -\log_2\sqrt{2}\sin(\theta/2). \tag{3.4b}$$

Thus for large n, $M^*(n,\theta)$ increases exponentially in n as $n \to \infty$ for $\theta < \pi/2$. As in the binary case an important unsolved problem is the determination of the "exponent" $R^*(\theta)$.

3.6. A "physical" interpretation of the Gaussian channel model. In this section we describe a channel model for which the input is a continuous-time function, and the noise is a sample from a white Gaussian noise process.

This discussion will not of course be mathematically rigorous because of the white noise assumption. A mathematically correct treatment of this subject can be found in [23].

Let us recall the model of the discrete-time Gaussian channel which we discussed above in § 3.1 and § 3.2. The input is an n-vector $\mathbf{x}$ which satisfies

$$\|\mathbf{x}\|^2 = Pn. \tag{3.5}$$

The output is an n-vector $\mathbf{y} = \mathbf{x} + \mathbf{z}$ where the components of $\mathbf{z}$ are independent, zero mean Gaussian variates with variance σ^2. Let us now assume that the channel can process γ numbers per second (we assumed above that $\gamma = 1$). Thus it will require $T = n/\gamma$ seconds to transmit $\mathbf{x}$. The channel capacity (maximum "error-free" rate) is then

$$C = \frac{\gamma}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) \quad \text{per second.}^{13} \tag{3.6}$$

Let us now consider a continuous-time channel for which the input is a real-valued function $x(t)$ that has support on a T second interval, say $[0, T]$. We assume

[13] That is, we can transmit roughly 2^{CT} messages reliably in T seconds when T is large.

that $x(t)$ satisfies the constraint

$$\frac{1}{T}\int_0^T x^2(t)\,dt = P_0, \tag{3.7}$$

so that P_0 is the average signal power. Let us expand $x(t)$ into a complex Fourier series on $[0, T]$,

$$x(t) = \frac{1}{T}\sum_{k=-\infty}^{\infty} x_k e^{ik2\pi/T}. \tag{3.8}$$

Let us assume that $x_0 = 0$. We say that $x(t)$ has *bandwidth* W cycles per second when $x_k = 0$, $|k| > WT$ (so that the highest frequency component of $x(t)$ is $2\pi W$ radians per second). Thus a signal with bandwidth W can be represented by an n-vector $\mathbf{x} = (x_{-n/2}, \cdots, x_{-1}, x_{+1}, \cdots, x_{n/2})$, where $n = 2[WT]$. Constraint (3.7) can be rewritten as

$$\|\mathbf{x}\|^2 = \sum_{\substack{k=-n/2 \\ k\neq 0}}^{n/2} x_k^2 = \left(\frac{P_0}{2W}\right)n. \tag{3.9}$$

The channel output is $y(t) = x(t) + z(t)$, where $z(t)$ is a sample from a (zero mean) white Gaussian noise process with (one-sided) spectral density N_0. We can formally write $y(t)$ in a Fourier series expansion on $[0, T]$,

$$y(t) = \frac{1}{T}\sum_{k=-\infty}^{\infty} y_k e^{ik2\pi t/T}, \tag{3.10a}$$

where

$$y_k = x_k + z_k, \tag{3.10b}$$

and the z_k are the coefficients of the corresponding expansion of $z(t)$. It is easily shown that the z_k are independent zero mean Gaussian variates with variance $N_0/2$.

Thus we see that this continuous-time model is mathematically identical with the discrete-time Gaussian channel with parameters $P = P_0/2W$, $\sigma^2 = N_0/2$, and $\gamma = 2W$. Thus from (3.6) its capacity is

$$C = W\log_2(1 + P_0/(N_0W)) \quad \text{per second}.$$

This is the celebrated formula given by Shannon in 1948 [1].

4. The general memoryless channel. In this section we shall define the general memoryless channel, the inputs and outputs of which belong to abstract sets. We shall then state the fundamental theorem which is of course a generalization of Theorem 2.1 and Theorem 3.1 for the binary symmetric and Gaussian channels respectively. Roughly speaking, this theorem states that corresponding to each channel there is a number called the "channel capacity;" and that when the channel is used in the system of Fig. 1, communication at rates less than channel capacity is possible with arbitrarily high reliability. Conversely, arbitrarily high reliability at rates exceeding capacity is not possible. It is felt that the reader who cares to digest the necessary preliminary definitions (§ 4.1), and the fundamental theorem

which follows (§ 4.2) will be amply rewarded with an insight into the brilliance of Shannon's contribution to the communication problem. Essentially this is the fact that an enormous class of seemingly vastly different communication channels can be unified by such a simply stated theorem.

In § 4.1 we give what is unfortunately a rather large collection of preliminary definitions. Since we are dealing with probabilistic notions about abstract sets, a measure theoretic approach is necessary. In § 4.2 we state the fundamental theorem for the general memoryless channel and give a proof of the direct half (asserting the existence of coding-decoding schemes with the required parameters).

The material in this section is essentially due to Shannon who proved the fundamental theorem for channels with finite input and output sets [1], [24] and for the Gaussian channel [17].[14] The generalization to abstract sets is due to Dobrusin [25] who used a slightly different approach. The treatment here which parallels Shannon's arguments closely does not seem to have appeared previously in the literature.

4.1. Definitions.

DEFINITION 1. Let $(X, \mathscr{A})$ and $(Y, \mathscr{B})$ be measurable spaces and let μ be a probability measure on the product space $(X \times Y, \mathscr{A} \times \mathscr{B})$, where $\mathscr{A} \times \mathscr{B}$ is the product σ-field. Let μ_X and μ_Y be the marginal probability measures induced by μ on X and Y respectively, and let $\mu_X \times \mu_Y$ be the product probability measure on the product space $(X \times Y, \mathscr{A} \times \mathscr{B})$. If μ is $(\mu_X \times \mu_Y)$-continuous (i.e., μ vanishes on $\mu_X \times \mu_Y$ null sets), denote the Radon–Nikodym derivative of μ with respect to $\mu_X \times \mu_Y$ by $d\mu/d(\mu_X \times \mu_Y) = d\mu/d\mu_X\, d\mu_Y$, and define the *information* of the measure μ on $X \times Y$ by

$$
\begin{aligned}
I(\mu) = I(\mu, X \times Y) &= E\left[\log_2 \frac{d\mu}{d\mu_X\, d\mu_Y}\right] \\
&= \int_{X \times Y} \left[\log_2 \frac{d\mu}{d\mu_X\, d\mu_Y}(x, y)\right] d\mu(x, y).
\end{aligned} \tag{4.1}
$$

The expectation in the right member of (4.1) is as indicated with respect to the measure μ. If μ is not $(\mu_X \times \mu_Y)$-continuous, the mutual information $I(\mu)$ is taken as $+\infty$. It is easy to show that $I(\mu) \geqq 0$, with equality if and only if $\mu = \mu_X \times \mu_Y$ (i.e., X, Y independent).

DEFINITION 2. *The general memoryless channel.* The channel *input* is a member x of X, and the channel *output* is a member y of Y. The channel is defined by a family $\{P_x, x \in X\}$ of ("transition") probability measures on Y. We interpret $P_x(B)$, $B \in \mathscr{B}$, as the conditional probability that the output belongs to B given that the input is x. We shall assume that for all $B \in \mathscr{B}$, $P_x(B)$ is a measurable function of x. Let μ_X be a probability distribution on the input set X. Then μ_X and $\{P_x, x \in X\}$ define a unique probability measure μ on $(X \times Y, \mathscr{A} \times \mathscr{B})$.[15] Corresponding to the input measure μ_X we can compute $I(\mu, X \times Y)$, the mutual

[14] In his 1948 paper [1], Shannon also considered, though in a less mathematically complete way, channels with real inputs and outputs.

[15] Let $S = A \times B$ $(A \in \mathscr{A}, B \in \mathscr{B})$; then $\mu(S) = \int_A P_x(B)\, d\mu_X(x)$, so that μ is defined for all sets of the form $A \times B$. It is easily shown that μ is a measure so that it extends uniquely to the product σ-field $\mathscr{A} \times \mathscr{B}$.

information of the measure μ on $X \times Y$. It is readily shown that

$$I(\mu, X \times Y) = \int_{X \times Y} \left[\log_2 \frac{d\mu}{d\mu_X \, d\mu_Y}(x, y) \right] d\mu(x, y) = \int_{X \times Y} \left[\log_2 \frac{dP_x}{d\mu_Y}(y) \right] d\mu(x, y). \tag{4.2}$$

There are two special cases for which (4.2) yields a useful computational formula for $I(\mu)$. The first is when X and Y are denumerable sets, so that all probability measures can be specified by giving the probability of each of the members of these sets. Thus we let $f_1(x) = \Pr\{X = x\}$, $f_2(y) = \Pr\{Y = y\}$, and $\hat{f}(y|x) = \Pr\{Y = y | X = x\}$. In this case $(dP_x|d\mu_Y)(y) = \hat{f}(y|x)/f_2(y)$ and (4.2) yields

$$I(\mu) = E \log_2 \frac{1}{f_2(y)} - E \log_2 \frac{1}{\hat{f}(y|x)} = \sum_{y \in Y} f_2(y) \log_2 \frac{1}{f_2(y)} - \sum_{x \in X} f_2(x) \sum_{y \in Y} \hat{f}(y|x) \log_2 \frac{1}{\hat{f}(y|x)}. \tag{4.3}$$

For example, if $X = Y = \{0, 1\}$ and $\hat{f}(y|x) = 1 - p_0$, $x = y$, and $\hat{f}(y|x) = p_0$, $x \neq y$, then the channel is the binary symmetric channel. Letting $f_1(0) = \pi_1$ and $f_1(1) = 1 - \pi_1$, we have from (4.3)

$$I(\mu) = f_2(0) \log_2 \frac{1}{f_2(0)} + f_2(1) \log_2 \frac{1}{f_2(1)} + p_0 \log_2 p_0 + (1 - p_0) \log_2 (1 - p_0), \tag{4.4}$$

where $f_2(0) = \pi_1(1 - p_0) + (1 - \pi_1)p_0$ and $f_2(1) = 1 - f_2(0)$.

The second special case is when $X, Y = R^n$, Euclidean n-space, and all probability measures are expressible by probability densities. Letting $f_1(x)$, $f_2(y)$ and $\hat{f}(y|x)$ be the densities for X, Y and the conditional density for Y given $X = x$ respectively, we have $(dP_x/d\mu_Y)(y) = \hat{f}(y|x)/f_2(y)$, and (4.2) yields

$$I(\mu) = \int_Y \left[\log_2 \frac{1}{f_2(y)} \right] f_2(y) \, dy + \int_X f_1(x) \, dx \int_Y \hat{f}(y|x) \, dy \left[\log_2 \frac{1}{\hat{f}(y|x)} \right]. \tag{4.5}$$

Of course (4.5) is meaningful only when these integrals are finite. For example if $X, Y = R$ and

$$\hat{f}(y|x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-y)^2/(2\sigma^2)},$$

then the channel is the Gaussian channel, and from (4.5) we have

$$I(\mu) = \int_{-\infty}^{\infty} \left[\log_2 \frac{1}{f_2(y)} \right] f_2(y) \, dy - \frac{1}{2} \log_2 2\pi e \sigma^2. \tag{4.6}$$

DEFINITION 3. Say we are given input and output spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ and a family of channel measures $\{P_x, x \in X\}$. Say we are also given a real-valued function Φ defined on X (which is $\mathcal{A}$-measurable). Let $\mathcal{M}$ be the set of measures

μ_X defined on the input space for which $E\Phi(X) \leqq 0$. ($\mathscr{M}$ may be the null set or the set of all measures on X.) We define the *channel capacity* C by

$$C = \sup_{\mu_X \in \mathscr{M}} I(\mu, X \times Y), \tag{4.7}$$

where μ is defined by P_x and μ_X. Note that although we have given the quantity in (4.17) the high sounding name of "channel capacity," we have not as yet given this quantity any real meaning.

Example 1. Let the channel be the binary symmetric channel and let $\Phi \equiv -1$, so that all measures on X belong to $\mathscr{M}$. Noting that for $0 \leqq \alpha \leqq 1$, $\alpha \log_2 (1/\alpha) + (1 - \alpha) \log_2 (1/(1 - \alpha)) \leqq 1$, with equality only at $\alpha = \frac{1}{2}$, we have from (4.4)

$$C = 1 + p_0 \log_2 p_0 + (1 - p_0) \log_2 (1 - p_0). \tag{4.8}$$

The equality $I(\mu) = C$ is attained when $\pi_1 = \frac{1}{2}$ so that $f_2(0) = f_2(1) = \frac{1}{2}$.

Example 2. Let the channel be the Gaussian channel with noise variance σ^2, and let $\Phi(x) = x^2 - P$. Thus a measure on the input space belongs to $\mathscr{M}$ if and only if $EX^2 \leqq P$. Now $E(Y^2|X = x) = \sigma^2 + x^2$ so that $EY^2 \leqq \sigma^2 + P$. Further it is easy to show that for any density $\alpha(y)$ for which $\int_{-\infty}^{\infty} y^2 \alpha(y)\, dy \leqq A$,

$$\int_{-\infty}^{\infty} \alpha(y) \log_2 \frac{1}{\alpha(y)}\, dy \leqq \frac{1}{2} \log_2 2\pi e A,$$

with equality if and only if $\alpha(y)$ is Gaussian with mean zero and variance A.

Thus from (4.6),

$$\begin{aligned} C &= \tfrac{1}{2} \log_2 2\pi e(P + \sigma^2) - \tfrac{1}{2} \log_2 2\pi e\sigma^2 \\ &= \tfrac{1}{2} \log_2 (1 + P/\sigma^2), \end{aligned} \tag{4.9}$$

where the equality $I(\mu) = C$ is attained when X is Gaussian with mean zero and variance P.

DEFINITION 4. Let $\{P_x, x \in X\}$ be a family of transition probability measures defining a channel with input space $(X, \mathscr{A})$ and output space $(Y, \mathscr{B})$. For $n = 1, 2, \cdots$, the *n-th extension of the channel* is a new channel with input space $(X^n, \mathscr{A}^n)$, output space $(Y^n, \mathscr{B}^n)$ and a family of probability measures $\{P_x^{(n)}, \mathbf{x} \in X^n\}$ on $(Y^n, \mathscr{B}^n)$. The "memoryless" aspect of the extension is expressed by the requirement that

$$P^{(n)}_{(x_1, x_2, \cdots, x_n)} = \prod_{k=1}^{n} P_{x_k}, \tag{4.10}$$

i.e., $P_{\mathbf{x}}^{(n)}$ is the product of the measures P_{x_k}, $k = 1, 2, \cdots, n$, on the coordinate output spaces. If $B \in \mathscr{B}^n$, we interpret $P_{\mathbf{x}}^{(n)}(B)$ as the conditional probability that the n-tuple $\mathbf{y} \in B$ given that the input is the n-tuple $\mathbf{x}$. Equation (4.10) tells us that the kth output y_k depends only on the kth input x_k, $k = 1, 2, \cdots, n$.

DEFINITION 5. Codes and error probability. Let $(X, \mathscr{A})$, $(Y, \mathscr{B})$, $\{P_x, x \in Y\}$ define a channel as in Definition 2 and let Φ be a real-valued function of X. Then a *code* with constraint corresponding to Φ is a set of M n-tuples $\mathbf{x}_i =$

$(x_{i1}, x_{i2}, \cdots, x_{in}) \in X^n$, $i = 1, 2, \cdots, M$, which satisfy

$$\sum_{k=1}^{n} \Phi(x_{ik}) \leqq 0. \tag{4.11}$$

For example, when the channel is the binary symmetric channel, $\Phi \equiv -1$, and a code is any set of M binary n-tuples. When the channel is the Gaussian channel taking $\Phi(x) = x^2 - P$ corresponds to constraint (3.1) with $A = nP$.[16]

A *decoding rule* is a ($\mathscr{B}^n$-measurable) function D which assigns to each n-tuple $\mathbf{y} \in Y^n$ an integer i, $1 \leqq i \leqq M$. The *transmission rate* R is defined by $R = (1/n) \log_2 M$, so that $M = 2^{nR}$.

If $\mathbf{x}_1$ is the input to the nth extension of the channel, an error is made if and only if the received n-tuple $\mathbf{y} \notin B_i \overset{\Delta}{=} \{\mathbf{y} : D(\mathbf{y}) = i\}$. Thus if $\mathbf{x}_i$ is the channel input, the resulting error probability is

$$P_{ei} = 1 - P_{\mathbf{x}_i}^{(n)}(B_i). \tag{4.12}$$

We assume that all M code members are equally likely to be transmitted, so that the overall error probability is

$$P_e = \frac{1}{M} \sum_{i=1}^{M} P_{ei}. \tag{4.13}$$

4.2 Fundamental theorem. With the above definitions in hand we state the following theorem.

FUNDAMENTAL THEOREM. *Let $(X, \mathscr{A})$, $(Y, \mathscr{B})$ and $\{P_x, x \in X\}$ define a channel, and let Φ be a function defining a coding constraint. Let the channel capacity be $C \leqq \infty$. Let $R < C$ be fixed. Then for any $\lambda > 0$, there exists an n (sufficiently large) and a code with $2^{[nR]}$ code words with error probability (on the n-th extension of the channel) $P_e \leqq \lambda$.*

Further, if $R > C$, there exists a $\lambda = \lambda(R) > 0$, such that any code with rate R has error probability $P_e \geqq \lambda(R)$.

We conclude with a proof of the direct-half of the fundamental theorem (asserting the existence of codes).

Proof of the direct half of the fundamental theorem with no coding constraint. In this section we give the proof of the theorem for the case $\Phi(x) \equiv -1$ so that there is no constraint on the code words. The extension to arbitrary Φ follows. The proof closely parallels that of the special case of the binary symmetric channel (Theorem 2.1) with another function called the "discrepancy" used instead of the Hamming distance.

Let $R < C$ be given. We shall study codes with dimension n and with $M = 2^{[Rn]}$ code words. Let $\tilde{\mu}_X$ be a measure on the input space X which induces a measure $\tilde{\mu}$ on the product space $X \times Y$ such that $I(\tilde{\mu}) > R$. Let $\tilde{\mu}_Y$ be the marginal distribution induced on the output space Y by $\tilde{\mu}$.[17] Let the measures $\tilde{\mu}_X^{(n)}$, $\tilde{\mu}_Y^{(n)}$, $\tilde{\mu}^{(n)}$ be the

[16] In § 3 we required that $\sum_1^n x_k^2 = Pn$. For large n this is almost equivalent to $\sum_1^n x_k^2 \leqq Pn$. This follows since if $\sum_1^n x_k^2 < Pn$, addition of an $(n+1)$th coordinate $x_{n+1} = [(n+1)P - \sum_1^n x_k^2]^{1/2}$ makes $\sum_1^{n+1} x_k^2 = (n+1)P$ and barely affects the rate.

[17] We also assume that the measure $\tilde{\mu}$ is $(\tilde{\mu}_X \times \tilde{\mu}_Y)$-continuous so that $I(\tilde{\mu})$ is given by (4.1). If it is not possible to find such a $\tilde{\mu}$ (in this case of course $C = \infty$), then another proof is needed (see Dobrusin [25]).

corresponding product measures on X^n, Y^n and $X^n \times Y^n$. It is not hard to show that for any measurable function $g(\mathbf{x}, \mathbf{y})$

$$\int_{X^{(n)}} d\tilde{\mu}_X^{(n)}(\mathbf{x}) \int_{Y^{(n)}} g(\mathbf{x}, \mathbf{y})\, dP_{\mathbf{x}}^{(n)}(\mathbf{y}) = \int_{X^n \times Y^n} g(\mathbf{x}, \mathbf{y})\, d\tilde{\mu}^{(n)}(\mathbf{x}, \mathbf{y}), \tag{4.14}$$

where $P_{\mathbf{x}}^{(n)}(\mathbf{y})$ is defined by (4.10).

Let $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in X^n$, $\mathbf{y} = (y_1, y_2, \cdots, y_n) \in Y^n$, and define the *discrepancy* between $\mathbf{x}$ and $\mathbf{y}$ by

$$d(\mathbf{x}, \mathbf{y}) = -\log_2 \frac{d\tilde{\mu}^{(n)}}{d\tilde{\mu}_X^{(n)}\, d\tilde{\mu}_Y^{(n)}}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} D(x_k, y_k), \tag{4.15a}$$

where

$$D(\mathbf{x}, \mathbf{y}) = -\log_2 \frac{d\tilde{\mu}(x, y)}{d\tilde{\mu}_X\, d\tilde{\mu}_Y}, \quad x \in X, \quad y \in Y. \tag{4.15b}$$

For arbitrary α, define

$$B(\alpha) = \{(\mathbf{x}, \mathbf{y}) \in X^n \times Y^n : d(\mathbf{x}, \mathbf{y}) \leqq \alpha n\},$$

and let $B^c(\alpha)$ be its complement. If follows from the weak law of large numbers that for $\alpha > I(\tilde{\mu})$,

$$\tilde{\mu}^{(n)}(B^c(\alpha)) \to 0 \quad \text{as } n \to \infty \tag{4.16}$$

(since $D(X, Y)$ has expectation $-I(\tilde{\mu})$ with respect to the measure $\tilde{\mu}$ on $X \times Y$).

We now proceed with our proof paralleling the proof of Theorem 2.1. Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M$ be a code with parameters M and n. We use the same decoding rule as in part (a) of the proof of Theorem 2.1 with discrepancy used instead of Hamming distance to produce an upper bound Q_ρ (ρ to be specified later) to the error probability:

$$P_e \leqq Q_\rho(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M) = \frac{1}{M} \sum_{k=1}^{M} f_1(\mathbf{x}_k) + \frac{1}{M} \sum_{k=1}^{M} \sum_{j \neq k} f_2(\mathbf{x}_k, \mathbf{x}_j), \tag{4.17a}$$

where

$$f_1(\mathbf{x}_k) = \int_{Y^n} (1 - \Phi_0(\mathbf{x}_k, \mathbf{y}))\, dP_{\mathbf{x}_k}(\mathbf{y}), \tag{4.17b}$$

$$f_2(\mathbf{x}_k, \mathbf{x}_j) = \int_{Y^n} \Phi_0(\mathbf{x}_j, \mathbf{y})\, dP_{\mathbf{x}_k}(\mathbf{y}) \tag{4.17c}$$

and

$$\Phi_0(\mathbf{x}, \mathbf{y}) = \begin{cases} 1, & d(\mathbf{x}, \mathbf{y}) \leqq \rho n, \\ 0, & d(\mathbf{x}, \mathbf{y}) > \rho n. \end{cases} \tag{4.17d}$$

As in Theorem 2.1 we now assume that the code words are chosen independently from X^n with identical measure $\tilde{\mu}_X^{(n)}$. Q_ρ, f_1 and f_2 are now random variables

with expectations $\bar{Q}_\rho, \bar{f}_1$ and $\bar{f}_2$ respectively. Consider first

$$\bar{f}_1 = \int_{X^n} d\tilde{\mu}_X^{(n)}(\mathbf{x}) \int_{Y^n} (1 - \Phi_0(\mathbf{x}, \mathbf{y}))\, dP_{\mathbf{x}}^{(n)}(\mathbf{y}). \tag{4.18}$$

From (4.14),

$$\bar{f}_1 = \int_{X^n \times Y^n} (1 - \Phi_0(\mathbf{x}, \mathbf{y}))\, d\tilde{\mu}^{(n)}(\mathbf{x}, \mathbf{y}) = \tilde{\mu}^{(n)}(B^c(\rho)). \tag{4.19}$$

Now consider

$$\bar{f}_2 = \int_{X^n} d\tilde{\mu}_X^{(n)}(\mathbf{x}_j) \int_{X^n} d\tilde{\mu}_X^{(n)}(\mathbf{x}_k) \int_{Y^n} \Phi_0(\mathbf{x}_j, \mathbf{y})\, dP_{\mathbf{x}k}^{(n)}(\mathbf{y}). \tag{4.20}$$

Again from (4.14) the two right most integrals in (4.20) are

$$\int_{X^n \times Y^n} \Phi_0(\mathbf{x}_j, \mathbf{y})\, d\tilde{\mu}^{(n)}(\mathbf{x}_k, \mathbf{y}),$$

where $\mathbf{x}_j$ is fixed. Thus this quantity equals

$$\int_{Y^n} \Phi_0(\mathbf{x}_j, \mathbf{y})\, d\tilde{\mu}_Y^{(n)}(\mathbf{y}),$$

Returning to (4.20) we have

$$\begin{aligned} \bar{f}_2 &= \int_{X^n \times Y^n} \Phi_0(\mathbf{x}, \mathbf{y})\, d\tilde{\mu}_X^{(n)}(\mathbf{x})\, d\tilde{\mu}_Y^{(n)}(\mathbf{y}) \\ &= \int_{B(\rho)} d\tilde{\mu}_X^{(n)}(\mathbf{x})\, d\tilde{\mu}_Y^{(n)}(\mathbf{y}). \end{aligned} \tag{4.21}$$

Now from (4.14) and (4.15) if $(\mathbf{x}, \mathbf{y}) \in B(\rho)$,

$$\frac{d\tilde{\mu}^{(n)}}{d\tilde{\mu}_X^{(n)}\, d\tilde{\mu}_Y^{(n)}}(\mathbf{x}, \mathbf{y}) \geqq 2^{-\rho n}. \tag{4.22}$$

Thus

$$\bar{f}_2 \leqq \int_{B(\rho)} 2^{\rho n}\, d\tilde{\mu}^{(n)}(\mathbf{x}, \mathbf{y}) \leqq 2^{\rho n}. \tag{4.23}$$

Returning to (4.17a) and using (4.19) and (4.23) we obtain

$$\bar{Q}_\rho \leqq \tilde{\mu}^{(n)}(B^c(\rho)) + M2^{-\rho n}. \tag{4.24}$$

Setting $\rho = -\frac{1}{2}(I(\tilde{\mu}) + R)$, we have from (4.16), $\tilde{\mu}^{(n)}(B^c(\rho)) \xrightarrow{n} 0$. The second term is $2^{-n/(2I(\tilde{\mu})-R))} \xrightarrow{n} 0$, so that we conclude that $\bar{Q}_\rho \xrightarrow{n} 0$, completing the proof.

Proof for channels with a coding constraint. In this section we modify our proof to account for an arbitrary constraint function Φ. Let $\tilde{\mu}_X$, $\tilde{\mu}_Y$ and $\tilde{\mu}$ be the measures in the above proof. Since $\tilde{\mu}_X \in \mathcal{M}$,

$$\gamma \triangleq \int \Phi(x)\, d\tilde{\mu}_X(x) \leqq 0. \tag{4.25}$$

Let us first assume that $\gamma < 0$. Let $X_1, X_2, \cdots, X_n$ be independent choices from X, each with measure $\tilde{\mu}_X$. Then the random n-vector $(X_1, X_2, \cdots, X_n) \in X^n$ has measure $\tilde{\mu}_X^{(n)}$. By the weak law of large numbers,

$$\tilde{\mu}_X^{(n)}\left\{\sum_{k=0}^{n} \Phi(x_k) \leqq 0\right\} \to 1 \quad \text{as } n \to \infty. \tag{4.26}$$

In the random experiment which we used in the above proof, we chose the M code words independently with probability measure $\tilde{\mu}_X^{(n)}$. Let $\mathscr{C}(= X^{nM})$ be the space of all possible codes, and let m be the measure on $\mathscr{C}$ which results from this random experiment. Letting $c = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M)$ be a code, we showed in the above proof that

$$\bar{Q} = \int_{\mathscr{C}} Q(c)\, dm(c) \to 0 \quad \text{as } n \to \infty. \tag{4.27}$$

where $Q = Q_\rho(c)$ is defined above. From this we deduced the existence of a code in $\mathscr{C}$ with $2^{[nR]}$ members and arbitrarily small error probability. If instead of averaging over $\mathscr{C}$ as in (4.27) we average over a subset of $\mathscr{C}$ and show that this average also tends to zero, we can deduce the existence of a code in the subset with the same vanishingly small error probability. Hence we define $\mathscr{G} \subseteq \mathscr{C}$ as the set of codes for which at least half the code words satisfy constraint (4.11). It follows from (4.26) that as $n \to \infty$,

$$m(\mathscr{G}) \to 1. \tag{4.28}$$

Also let us define the measure $\hat{m}$ on $\mathscr{G}$ by

$$\hat{m}(G) = \frac{m(G)}{m(\mathscr{G})}, \qquad G \subseteq \mathscr{G}. \tag{4.29}$$

In other words, $\hat{m}(G)$ is the conditional probability of G given $\mathscr{G}$. Further,

$$\begin{aligned} \hat{Q} &\triangleq \int_{c \in \mathscr{G}} Q(c)\, d\hat{m}(c) = \frac{1}{m(\mathscr{G})} \int_{c \in \mathscr{G}} Q(c)\, dm(c) \\ &= \frac{\bar{Q}}{m(\mathscr{G})} \to 0 \quad \text{as } n \to \infty. \end{aligned} \tag{4.30}$$

Thus we are guaranteed the existence of a code c^* in $\mathscr{G}$ with the appropriate number of code words ($2^{[nR]}$) and with arbitrarily small error probability. Excluding from c^* those code words which do not satisfy constraint (4.11) (no more than half) does not increase P_e and leaves the rate R essentially unchanged for large n. This completes the proof for the case $\gamma = \int \Phi\, d\tilde{\mu}_X < 0$. The proof for the case $\gamma = 0$ follows from the following lemma which asserts the existence of another measure $\tilde{\mu}'_X$ on X for which $\int \Phi\, d\tilde{\mu}'_X < 0$, and which induces a measure $\hat{\mu}'$ on $X \times Y$ with $I(\tilde{\mu}')$ arbitrarily close to $I(\tilde{\mu})$.

LEMMA. *Let μ_X be a measure on X for which*

(i) $\Phi(x) \neq 0$ *on a set of positive measure,*

(ii) $\int_X \Phi(x)\, d\mu_X = 0$,

(iii) $I(\mu) = I_0$, *where μ is induced on $X \times Y$ by μ_X.*

Then for any $\gamma > 0$, there exists a measure μ'_X on X such that

(iv) $\int_X \Phi(x)\, d\mu'_X(x) < 0$,

(v) $I(\mu') \geqq I_0 - \gamma$, *where μ' is induced on $X \times Y$ by μ'_X.*

Proof. Let $X_+ = \{x \in X : \varphi(x) \geqq 0\}$ and $X_- = \{x \in X : \varphi(x) < 0\}$. Let A be any measurable subset of X, and define

$$\mu'_X(A) = \frac{1+\varepsilon}{1+\kappa\varepsilon}\mu_X(AX_-) + \frac{(1-\varepsilon)}{1+\kappa\varepsilon}\mu_X(AX_+), \tag{4.31}$$

where $\kappa = \mu_X(X_-) - \mu_X(X_+)$, and ε, $0 < \varepsilon < 1$, is arbitrary. Clearly (iv) is satisfied. With the measures μ and μ_Y, μ' and μ'_Y defined in the usual way we show that $I(\mu') \geqq I_0 - \delta(\varepsilon)$, where $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$. The lemma then follows.

For any Y-measurable set B,

$$\begin{aligned} \mu'_Y(B) &= \int_X P_x(B)\, d\mu'_X(x) \\ &= \int_{X_+} \frac{1-\varepsilon}{1+\kappa\varepsilon} P_x(B)\, d\mu_X(x) + \int_{X_-} \frac{1+\varepsilon}{1+\kappa\varepsilon} P_x(B)\, d\mu_X(x) \\ &\leqq \frac{1+\varepsilon}{1+\kappa\varepsilon}\mu_Y(B). \end{aligned} \tag{4.32}$$

Thus for almost all x, y,

$$\frac{dP_x(y)}{d\mu'_Y} \geqq \frac{1+\kappa\varepsilon}{1+\varepsilon}\frac{dP_x(y)}{d\mu_y}. \tag{4.33}$$

Hence

$$\begin{aligned} I(\mu') &= \int_X d\mu'_X(x) \int_Y dP_x(y) \left(\log_2 \frac{dP_x(y)}{d\mu'_Y} \right) \\ &\geqq \int_X d\mu'_X(x) \int_Y dP_x(y) \left(\log_2 \frac{dP_x(y)}{d\mu_Y} \right) + \log_2 \frac{1+\kappa\varepsilon}{1+\varepsilon}. \end{aligned} \tag{4.34}$$

Finally, from (4.31) we have

$$I(\mu') \geqq \frac{1-\varepsilon}{1+\kappa\varepsilon} I(\mu) + \log_2 \frac{1+\kappa\varepsilon}{1+\varepsilon} \to I(\mu) \quad \text{as } \varepsilon \to 0,$$

completing the proof.

REFERENCES

[1] C. E. SHANNON, *A mathematical theory of communication*, Bell System Tech. J., 27 (1948), pp. 379–423, 623–656. Reprinted in C. E. SHANNON AND W. WEAVER, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949.

[2] R. G. GALLAGER, *Information Theory and Reliable Communication*, John Wiley, New York, 1968.

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

[4] R. B. ASH, *Information Theory*, Interscience, New York, 1965.

[5] N. M. ABRAMSON, *Information Theory and Coding*, McGraw-Hill, New York, 1963.

[6] S. KOTZ, *Recent Results in Information Theory*, Methuen, London, 1966.

[7] R. G. GALLAGER, *A simple derivation of the coding theorem and some applications*, IEEE Trans. Information Theory, IT-11 (1965), pp. 3–17.

[8] C. E. SHANNON, R. G. GALLAGER AND E. R. BERLEKAMP, *Lower bounds to error probability for coding on discrete memoryless channels*, Information and Control, 10 (1968), pp. 65–103, 522–552.

[9] W. FELLER, *Introduction to Probability Theory*, vol. I, John Wiley, New York, 1950.

[10] R. W. HAMMING, *Error detecting and correcting codes*, Bell System Tech. J., 29 (1950), pp. 147–160.

[11] W. W. PETERSON, *Error Correcting Codes*, MIT Press, Cambridge, Massachusetts, 1961.

[12] E. R. BERLEKAMP, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

[13] E. N. GILBERT, *A comparison of signaling alphabets*, Bell System Tech. J., 31 (1952), pp. 504–522.

[14] M. PLOTKIN, *Binary codes of specified minimum distance*, IEEE Trans. Information Theory, IT-6 (1960), pp. 445–450.

[15] P. ELIAS, Unpublished work. Elias' result is given in [8, Theorem 7].

[16] A. D. WYNER, *Capabilities of bounded discrepancy decoding*, Bell System Tech. J., 44 (1965), pp. 1061–1122.

[17] C. E. SHANNON, *Communication in the presence of noise*, Proc. IRE, 37 (1959), pp. 10–21.

[18] ———, *Probability of error for optimal codes in the Gaussian channel*, Bell System Tech. J., 38 (1959), pp. 611–656.

[19] H. LANDAU AND D. SLEPIAN, *On the optimality of the regular simplex code*, Ibid., 45 (1966), pp. 1247–1272.

[20] A. H. NUTTAL, *Error probabilities for equicorrelated M-ary signals under phase-coherent and phase-incoherent reception*, IEEE Trans. Information Theory, IT-8 (1962), pp. 305–314.

[21] C. E. WEBER, *New solutions to the signal design problem for coherent channels*, Ibid., IT-12 (1966), pp. 161–167.

[22] R. A. RANKIN, *The closest packing of spherical caps in n-dimensions*, Proc. Glasgow Math. Assoc., 2 (1955), pp. 139–144.

[23] A. D. WYNER, *Capacity of the band-limited Gaussian channel*, Bell System Tech. J., 25 (1966), pp. 359–395.

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

[25] R. L. DOBRUSIN, *General formulation of Shannon's main theorem*, Uspehi Mat. Nauk., 14 (1959), pp. 3–104. Reprinted in Amer. Math. Soc. Transl. (2), 33 (1963), pp. 323–438.