1955 National Convention

IRE Convention Record

Klein

Part 4—
Computers, Information Theory, Automatic Control

SESSIONS ON . . .

- Automatic Control I
- Information Theory I
- Automatic Control II—Trends in Automatization of Procedures and Processes in Business and Industry
- Electronic Computers I
- Electronic Computers II — Symposium: Design of Machines to Simulate the Behavior of the Human Brain
- Information Theory II
- Electronic Computers III
- Information Theory III

SPONSORED BY
IRE PROFESSIONAL GROUPS ON . . .

- Automatic Control
- Electronic Computers
- Information Theory

Presented at the IRE National Convention, New York, N.Y., March 21-24, 1955

The Institute of Radio Engineers

CODING FOR NOISY CHANNELS*

Peter Elias
Department of Electrical Engineering and Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Summary: Shannon's and Feinstein's versions of the channel capacity theorem, specialized to the binary symmetric channel, are presented. A much stronger version is proved for this channel. It is shown that the error probability as a function of delay is bounded above and below by exponentials, whose exponents agree for a considerable range of values of the channel and the code parameters. In this range the average behavior of all codes is essentially optimum, but for small transmission rates this is not true. The results of this analysis are shown to apply to check-symbol codes of four kinds which have progressively simpler coding procedures. The last of these is error-free, and makes it possible to transmit information at a rate equal to the channel capacity with a probability one that no decoded symbol will be in error.

Introduction

Since Shannon[1, 2] showed that information could be transmitted over a noisy channel at a positive rate with an arbitrarily low probability of error at the receiver, there has been considerable interest in constructing specific transmission schemes that exhibit such behavior.

For a signal transmitted over a channel perturbed by Gaussian noise, Golay[3] and Fano[4] found schemes which in the limit had the desired behavior, but it was a limit of infinite bandwidth or vanishing transmission rate. Rice[5] investigated the characteristics of transmission using randomly selected noise waveforms, and got an indication of exponential decrease in error probability with increasing time delay. Feinstein[6] showed that the same sort of behavior, at least as an upper bound, held true for more general channels.

For the binary channel, Hamming[7], Gilbert[8], Plotkin[9], and Golay[10] investigated a variety of codes, and found some basic properties of the binary symmetric channel. Laemmel[11], Muller[12], and Reed[13] also constructed specific codes and classes of codes. The first constructive code for transmission at a nonzero rate over a noisy binary channel was discovered recently by the author[14]. The investigation reported in the present paper started as a continuation of that work, and an investigation of the rate at which the error probability decreased with delay originally developed from a comparison of check-symbol codes with codes of less restricted types. It seems more sensible to present the results in reverse order. After a definition of the channel and general coding procedures, Shannon's and Feinstein's channel capacity theorems are stated, and a stronger theorem is given for the binary symmetric channel, which shows in considerable detail the behavior of error probability at the receiver as a function of the parameters of the channel and the code, and the delay time. It is then shown that most of these results carry over to a variety of kinds of check-symbol codes. One of these, of primarily academic interest, is error-free[14], and permits the transmission of an infinite sequence of message symbols at an average rate equal to the channel capacity with a probablity one that no decoded digit is in error.

The Channel

The coding problem we will discuss is illustrated in Fig. 1. The problem is to match the output of an ideal binary message source to a binary symmetric noisy channel.

The message source generates a sequence of binary symbols, say the binary digits zero and one. Zeros and ones are selected with equal probability, and successive selections are statistically independent.

The channel accepts binary symbols as an input and produces binary symbols as an output. Each input symbol has a probability $p_o < 1/2$ of being received in error, and a probability $q_o = 1 - p_o$ of being received as transmitted. The transmission error probability p_o is a constant, independent of the value of the symbol being transmitted: the channel is as likely to turn a one into a zero as to turn a zero into a one. The channel, in effect, adds a noise sequence to the input sequence to produce the output sequence; the noise is a random sequence of zeros and ones, synchronous with the signal sequence, in which the ones have probability p_o, and the addition is addition modulo two of each signal digit to the corresponding noise digit $(1+1 = 0+0 = 0,\ 0+1 = 1+0 = 1)$.

If the message source were connected directly to the channel, a fraction p_o of the received symbols would be in error. A coding procedure for reducing the effect of the errors is shown in Figs. 1 and 2. The output of the message source is segmented into consecutive blocks of length M. There are 2^M such blocks, and they are selected by the source with equal probability. To each input block of M binary symbols is assigned an output block of N binary symbols, $N > M$.

The input sequences of length M are the messages to be sent; the output sequences of length N are the transmitted signals, and the correspondence between input and output blocks is the code used. The use of the word "code" is justified

*This work was supported in part by the Signal Corps; the Office of Scientific Research, Air Research and Development Command; and the Office of Naval Research.

by Fig. 2, where the correspondence between input and output blocks is given in the form of a codebook. On the left is a column of the 2^M possible messages, listed as M-digit binary numbers in numerical order. Following each message is the N-digit binary number which is the corresponding signal, so that the codebook has 2^M entries, each of which lists a message and the corresponding signal.

The system in operation is shown in Fig. 1. The source selects a message that is coded into a transmitted signal and sent over the noisy channel. The received block of N -- the received, or noisy, signal -- differs from the transmitted signal in about p_oN of its N symbol values. The decoder receives this noisy signal and reproduces one of the 2^M possible messages, with an average probability P_e of making an incorrect choice.

The most general type of decoder is shown in Fig. 3. It is a codebook with 2^N entries, one for each of the possible received signals. The left column is the received sequence, arranged as an N-digit binary number in numerical order. This is followed by the M-symbol message block that will be reproduced when that sequence is received as a noisy signal.

In order to minimize P_e, the codebook must be so constructed that the message that is selected when a given noisy signal is received is the one corresponding to the signal most likely to have been transmitted. For the binary symmetric channel, the signal most likely to have been transmitted is the one that differs from the received signal in the smallest number of symbol positions. This follows from the fact that a particular group of k errors has probability $p_o^k q_o^{N-k}$ of being introduced by the channel; this probability decreases as k increases, for $p_o < 1/2$.

For this channel, the codebook may be simplified. In fact, the transmitter codebook may be used in reverse order. The noisy signal is compared with each of the possible transmitted signals, and the number of positions in which they differ is counted. The signal with the lowest count is assumed to have been transmitted, and the corresponding message block is reproduced as the best guess at the transmitted message. This guess may, of course, be incorrect, and will be if the noise has altered more than half of the positions in which the transmitted signal differs from some other listed signal.

This decoding procedure may be described in a geometrical language introduced by Hamming[7]. Each signal is taken as a point or a vector in an N-dimensional space, with coordinates equal to the values (zero or one) of its N binary symbols. The distance between two points is defined as the number of coordinates in which they differ. In this language, the noisy signal is decoded as the nearest of the signal points, and the corresponding message is chosen.

For given M and N, the error probability P_e depends on the set of points that are used as signals. If these are clustered in a small part of the space, P_e will be large; if they are far from one another, P_e will be small. As specialized to this channel, Shannon's second coding theorem states an asymptotic relationship between M, N, and P_e for a suitable selection of signal points.

Channel Capacity and Error Probability

First, some definitions are required. Given a binary symmetric channel with transmission error probability p_o and $q_o = 1 - p_o$, the equivocation $E_o = E(p_o)$ and the capacity $C_o = C(p_o)$ of the channel, both measured in bits per symbol, are given by

$$\begin{aligned} E_o &= -p_o \log p_o - q_o \log q_o \\ C_o &= 1 - E_o \end{aligned} \tag{1}$$

(Here and later, all logarithms are to the base two.)

Given a coding procedure like that illustrated by Figs. 1, 2, and 3, the redundancy E_1 and the transmission rate C_1, also in bits per symbols, are given by

$$\begin{aligned} E_1 &= \frac{N - M}{N} \\ C_1 &= 1 - E_1 = \frac{M}{N} \end{aligned} \tag{2}$$

It is convenient to introduce the probability p_1 which is the upper bound of the transmission error probabilities for which this particular code can be expected to work, and $q_1 = 1 - p_1$. These are uniquely defined by

$$\begin{aligned} &p_1 < \tfrac{1}{2} \\ &E_1 = E(p_1) = -p_1 \log p_1 - q_1 \log q_1 \end{aligned} \tag{3}$$

since a plot of E(p) or C(p) is monotonic for $0 \leqslant p \leqslant 1/2$.

Finally, the average probability of an error in decoding, which was written as P_e above, will in general be a function of the block length N, the channel capacity C_o or error probability p_o, and the transmission rate C_1 or the probability p_1. It will be written as $P_e(N, p_o, p_1)$.

Shannon's second coding theorem[1], as applied to this channel, follows.

Theorem 1. Given any fixed $C_1 < C_o$ and any fixed $\epsilon > 0$, for all sufficiently large N there are codes which will transmit information at the rate C_1 bits per symbol and will decode it with an error probability per block of N, $P_e(N, p_o, p_1) < \epsilon$. This cannot be done for $C_1 > C_o$.

Shannon's proof of the theorem proves more than the theorem states. A code is a selection of

2^{NC_1} signal sequences from among 2^N possibilities. Including those codes that select the same signal two or more times to represent several different messages, there are $2^{N\cdot 2^{NC_1}}$ different codes. Each of these will have an average decoding error probability (averaged over the different messages, with equal weights). Shannon shows that the average of all of these (averaged over the different codes, with equal weights) is less than ϵ. Since the error probability for each code is positive, it follows that at least one code has an average error probability less than ϵ; and it also follows, as Shannon remarks, that, at most, a fraction f of the codes can have an average error probability as great as ϵ/f, so that almost all of the codes have arbitrarily small error probability; that is, almost all codes are "good" codes, although some "bad" codes do exist. By the same argument, in any one good code the error probability for most of the individual messages is less than ϵ/f, so that by discarding a few of the signal sequences and transmitting at a very slightly slower rate, any good code can be made into a uniformly good code. This result has considerable practical importance, since a uniformly good code will transmit with the specified small error probability, regardless of the probabilities with which message sequences are selected, and there are many information sources whose statistics are not known in detail.

The major question left open by this theorem is how large N must be for given p_o, p_1, and ϵ. Feinstein[6] has proved a stronger version which provides an upper bound for $P_e(N, p_o, p_1)$. As specialized to the binary symmetric channel it may be written as:

Theorem 2. Given any $C_1 < C_o$, an $\epsilon(p_o, p_1) > 0$ can be found. For any sufficiently large N, a code may be constructed which will transmit information at the rate C_1 bits per symbol which can be decoded with $P_e(N, p_o, p_1) < 2^{-\epsilon N}$.

Feinstein's proof consists of the construction of a code that satisfies the requirements of the theorem and is uniform in the sense that all signals are good signals. Some indication of the relation of ϵ to the channel and code parameters is also given.

The next theorem is much stronger than this, but unlike Shannon's and Feinstein's it does not apply to the general discrete noisy channel without memory, but only to the binary symmetric case. Some more definitions are needed. It turns out that the error probability P_e is bounded not only above but below by exponentials in N, and that for a considerable range of channel and code parameters the exponents of the two bounds agree. The error exponent for the best possible code is defined as

$$\alpha_{\text{opt}}(N, p_o, p_1) = \frac{-\log P_e(N, p_o, p_1)}{N} \qquad (4)$$

and $\alpha_{\text{avg}}(N, p_o, p_1)$ is defined as the same function of the average of the error probabilities of all codes.

An additional probability value is also needed, along with the values of α, C, and E which go with it:

$$\begin{aligned} p_{\text{crit}} &= \frac{p^{1/2}}{p^{1/2} + q^{1/2}}, \quad q_{\text{crit}} = 1 - p_{\text{crit}} \\ E_{\text{crit}} &= E(p_{\text{crit}}), \quad C_{\text{crit}} = 1 - E_{\text{crit}} \\ \alpha_{\text{crit}} &= \lim_{N\to\infty} \alpha_{\text{opt}}(N, p_o, p_{\text{crit}}) \end{aligned} \qquad (5)$$

Finally, the margin in error probability and the margin in channel capacity need labeling:

$$\begin{aligned} \delta &= p_1 - p_o \\ \Delta &= C_o - C_1 \end{aligned} \qquad (6)$$

For a binary symmetric channel with capacity C_o and transmission rate C_1, the following statements hold.

Theorem 3. (a) For $p_o < p_1 < p_{\text{crit}}$, $C_o > C_1 > C_{\text{crit}}$, the average code is essentially as good as the best code:

$$\begin{aligned} \alpha(p_o, p_1) &= \lim_{N\to\infty} \alpha_{\text{opt}}(N, p_o, p_1) = \lim_{N\to\infty} \alpha_{\text{avg}}(N, p_o, p_1) \\ &= -\Delta - \delta \log \frac{p_o}{q_o} \end{aligned} \qquad (7)$$

(b) For $p_{\text{crit}} < p_1 < 1/2$, the average code is not necessarily optimum; for p_1 near 1/2 it is certainly not. Specifically,

$$\begin{aligned} \alpha_{\text{avg}}(p_o, p_1) &= \lim_{N\to\infty} \alpha_{\text{avg}}(N, p_o, p_1) \\ &= \alpha_{\text{crit}} + C_{\text{crit}} - C_1 \end{aligned} \qquad (8)$$

where α_{crit} is the $\alpha(p_o, p_1)$ of Eq. (5) with $p_1 = p_{\text{crit}}$, while for α_{opt} there are two upper and two lower bounds:

$$\liminf \alpha_{\text{opt}}(N, p_o, p_1) \geqslant \begin{cases} \alpha_{\text{crit}} + C_{\text{crit}} - C_1 \\ \frac{p_1}{2} \log \frac{1}{4pq} - C_1 \end{cases} \qquad (9)$$

$$\limsup \alpha_{\text{opt}}(N, p_o, p_1) \leqslant \begin{cases} -\Delta - \delta \log \frac{p_o}{q_o} \\ \frac{E_1}{4} \log \frac{1}{4pq} \end{cases} \qquad (10)$$

As $p_1 \to 1/2$, the second bound in (9) approaches the second bound in (10);

$$\lim_{N\to\infty} \lim_{p_1 \to 1/2} \alpha_{\text{opt}}(N, p_o, p_1) = \frac{1}{4} \log \frac{1}{4pq} \tag{11}$$

which is always greater than

$$\alpha_{\text{avg}}\left(p_o, \frac{1}{2}\right) = \alpha_{\text{crit}} + C_{\text{crit}} \tag{12}$$

The content of this theorem is illustrated by Fig. 4. This is a plot of the channel capacity $C(p)$ vs. transmission error probability p for a binary symmetric channel. A dashed line is drawn tangent to the curve at the point given by the channel parameters p_o, C_o. This tangent line has the slope $\log (p_o/q_o)$. The critical point p_{crit}, C_{crit} is the point at which the slope of the curve is $(1/2) \log (p_o/q_o)$. For $p_o < p_1 < p_{\text{crit}}$, the $\alpha(p_o, p_1)$ of (7), which is both the average and the optimum error exponent, is the length of a vertical dropped from the channel capacity curve to the tangent line at the ordinate p_1.

At $p_1 = p_{\text{crit}}$, the dotted line that determines $\alpha_{\text{avg}}(p_o, p_1)$ diverges from the tangent line. For $p_{\text{crit}} < p_1 < 1/2$ the exact value of $\alpha_{\text{opt}}(N, p_o, p_1)$ is not known, but is given by the length of a vertical at ordinate p_1, dropped from the channel capacity curve and terminating in the shaded region. The upper and lower bounds of this region provide lower and upper bounds, respectively, on the value of α_{opt}. These bounds converge to $(1/4) \log (1/4pq)$ at $p_1 = 1/2$, and near this point α_{opt} is definitely $> \alpha_{\text{avg}}$.

The value of α given by the tangent line at $p_1 = 1/2$, although not approached for the transmission of information at any nonzero rate, is the correct value of α_{opt} for transmission of one bit per block of N symbols.

An outline of the proof of Theorem 3 appears in the Appendix. A more detailed presentation, giving bounds on $P_e(N, p_o, p_1)$, as well as on α, will appear elsewhere.

Check-Symbol Codes

The preceding three theorems are interesting in theory but discouraging in practice. They imply that a good code will require a transmitting codebook containing $N \cdot 2^{NC_1}$ binary digits in all, and either a receiver codebook containing $N \cdot 2^N$ binary digits or another copy of the transmitter codebook and 2^{NC_1} comparisons of the received signal with the possible transmitted signals. Since in interesting cases NC_1 may be of the order of 100, the requirements in time and space are unmanageable. Furthermore, it would be quite consistent with these theorems if no code with any simplicity or symmetry properties were a good code.

The theorems that follow show that this is fortunately not the case. Four kinds of codes of increasing simplicity and convenience from the point of view of realization are demonstrated to have essentially the same behavior, from both a channel capacity and an error probability point of view, as the optimum code. The last of the four is of theoretical interest as well, since it permits the receiver to set the decoding error probability arbitrarily low without consulting the transmitter.

A check-symbol code of block length N is a code in which the 2^{NC_1} signal sequences have in their initial NC_1 positions all 2^{NC_1} possible combinations of symbol values. The first NC_1 positions will be called information positions and the last NE_1 will be called check positions. The signal corresponding to a message sequence is that one of the signal sequences whose initial symbols are the message.

A parity check-symbol (pcs) code is a check-symbol code in which the check positions are filled in with digits each of which completes a parity check of some of the information positions. Such codes were discussed in detail first by Hamming[1], who calls them systematic codes. A pcs code is specified by an $NC_1 \times NE_1$ matrix of zeros and ones, the ones in a row giving the locations of the information symbols whose sum modulo two is the check digit corresponding to that row. The process is illustrated in Fig. 5. Such a code requires $NC_1 \times NE_1 = N^2 C_1 E_1 \leqslant \frac{1}{4} N^2$ binary digits in its codebook, these being the digits in the check-sum matrix.

A sliding pcs code is defined as a pcs code in which the check-sum matrix is constructed from a sequence of N binary symbols by using the first NC_1 of them for the first row, the second to $(NC_1 + 1)$st for the second row ..., the NE_1th to the Nth for the NE_1th row. This code requires only an N-binary-digit codebook.

Finally a convolutional pcs code is defined as one in which check symbols are interspersed with information symbols, and the check symbols check a fixed pattern of the preceding NC_1 information positions if $C_1 \geqslant 1/2$; if $E_1 > 1/2$, the information symbols add a fixed pattern of zeros and ones to the succeeding NE_1 check positions. Such a code requires $\max(NC_1, NE_1) \leqslant N$ binary digits in its codebook. It is illustrated by Fig. 6.

Theorem 4. All the results of Theorem 3 apply to check-symbol codes and to pcs codes. The results of part (a) of that theorem apply to sliding pcs codes.

In reading Theorem 3 into Theorem 4, the average involved in α_{avg} is the average of all codes of the appropriate type; that is, all combinations of check symbols for the check-symbol codes, all check-sum matrices for the pcs codes, all sequences of N binary digits for the sliding pcs code.

Theorem 5. The results of part (a) of

Theorem 3 apply to convolutional pcs codes, if $P_e(N, p_o, p_1)$ is interpreted as the error probability per decoded symbol. For infinite memory (each check symbol checking a set of prior information symbols extending back to the start of transmission over the channel) the N in $P_e(N, p_o, p_1)$ for a particular decoded information symbol is the number of symbols which have been received since it was received.

This theorem shows that error-free coding can be attained at no loss either in channel capacity or in error probability, a question raised by the author when the first error-free code was introduced[14]. By waiting long enough, the receiver can obtain as low a probability of error per digit as is desired, without a change of code being necessary. By gradually reducing the ratio of check to information symbols toward E_o/C_o, using the law of the iterated logarithm for binary sequences, it can be shown that in an infinite sequence of message digits transmission is obtained at average rate C_o with probability one of no errors in the decoded message.

Conclusion

An appreciable gain in simplicity has been achieved in going from an arbitrary average code to a convolutional or sliding pcs code. It is possible to encode and decode either of these codes with a codebook of only N or fewer binary digits. However, the decoding operation will require 2^{NC_1} or 2^{NE_1} (whichever is smaller) comparisons, which will take a great deal of time in interesting cases. No decoding procedure that replaces this operation by a small amount of computing has yet been discovered, although the iterated Hamming code, which is error-free[14], gives hope that it may be possible to manage this while still keeping optimum behavior in terms of channel capacity and error probability -- a feature which the iterated Hamming code lacks.

Acknowledgments

After the analytical work reported in this paper was done, but before it had been organized for presentation, I discovered that Dr. Shannon was also working on the problem of error probability, and was to present his results at the same meeting. In discussing the results with Dr. Shannon, he mentioned the geometric relationship between the tangent line and the capacity curve, illustrated in Fig. 4, in the region $p_1 < p_{crit}$. I do not know whether this would have occurred to me in organizing my results, but I do know that it is vital; the information to the right of p_{crit} is my own, but is impossible to present in any other fashion without getting lost in numbers of families of curves.

It is a pleasure to acknowledge my indebtedness to the atmosphere at the Research Laboratory of Electronics, without which this work would not have gotten started; and to my colleagues, Professors Fano, Huffman, and Yngve, who provided that part of the atmosphere most relevent to this specific project.

Appendix

1. Outline Proof of Theorem 3

Using the symbols and definitions of Eqs. (1), (3), (5), and (6), let $k_1 = Np_1$ be an integer. Define $V_N(k)$, the volume of an N-dimensional sphere of radius k, by

$$V_N(k) = \sum_{j=0}^{k} \binom{N}{j} = \sum_{j=0}^{k} \frac{N!}{j!\,(N-j)!} \qquad (A.1)$$

Select $2^N/V_N(k_1)$ sequences as signaling sequences. Then the signaling rate is

$$\frac{1}{N} \log \left\{2^N/V_N(k_1)\right\} = 1 - \frac{1}{N} \log V_N(k_1) \qquad (A.2)$$

If the selection of signal sequences can be made so that every possible received sequence differs from one (and only one) signal sequence in k_1 or fewer positions, then the probability of a detection error will be just the probability $P_I(N, p_o, p_1)$ that more than k_1 out of N errors are made in transmission. This is the tail of the binomial distribution:

$$P_I(N, p_o, p_1) = \sum_{j=k_1+1}^{N} p_o^j\, q_o^{N-j} \binom{N}{j} \qquad (A.3)$$

Such a selection is not, in general, possible. However, $P_I(N, p_o, p_1)$ of (A.3) is a lower bound to the average decoding error probability $P_e(N, p_o, p_1)$ for any actual selection of signal points: this follows directly from the fact that $p_o^j\, q_o^{N-j}$ is a monotonically decreasing function of j.

The average of all possible codes is used to provide an upper bound to the decoding error probability of the best code. The average probability of a detection error, $P_{III}(N, p_o, p_1)$, is the probability $P_{II}(N, j, k_1)$ of a decoding error when just j transmission errors have occurred, averaged over the binomial distribution of j. With Eq. (A.3) this gives

$$\begin{aligned} P_{III}(N,p_o,p_1) &= \sum_{j=0}^{N} P_{II}(N,j,k_1)\, p_o^j\, q_o^{N-j} \binom{N}{j} \\ &\leq \sum_{j=0}^{k_1} P_{II}(N,j,k_1)\, p_o^j\, q_o^{N-j} \binom{N}{j} \\ &\qquad + P_I(N,p_o,p_1) \end{aligned} \qquad (A.4)$$

The probability $P_{II}(N, j, k_1)$ of a decoding error when just j transmission errors have occurred is the probability that one of the $\{2^N/V_N(k_1)\} - 1$ incorrect signal sequences differs in j or fewer places from the received sequence. There are a total of $V_N(j)$ sequences which differ from the received sequence in as few as j positions, and the probability of missing all of them in $\{2^N/V_N(k_1)\} - 1$ tries is, for $j < k_1$, bounded by

$$\left(1 - \frac{V_N(j)}{2^N}\right)^{\{2^N/V_N(k_1)\}-1} \geqslant 1 - \frac{V_N(j)}{V_N(k_1)} \qquad (A.5)$$

Equation (A.5) gives the probability of no decoding error: P_{II} is the probability of a decoding error, so

$$P_{II}(N, j, k_1) \leqslant \frac{V_N(j)}{V_N(k_1)} \leqslant \frac{\binom{N}{j}}{\binom{N}{k_1}} \qquad (A.6)$$

Equations (A.4) and (A.6) give

$$P_{III}(N, p_o, p_1) \leqslant \sum_{j=0}^{k_1} p_o^j q_o^{N-j} \frac{\binom{N}{j}^2}{\binom{N}{k_1}} + P_I(N, p_o, p_1) \qquad (A.7)$$

Now the sums in Eqs. (A.1) and (A.3) are bounded below by the value of their largest term and above by a geometric series multiplied by that term -- the last term in Eq. (A.1), the first in Eq. (A.3). (See Feller[15], p. 126 for the bounds for Eq. (A.3).) The sum in Eq. (A.7) is similarly bounded above and below, if p_1 is less than p_{crit}, which is the condition guaranteeing that the last term in the sum is the largest. Using these results, taking logarithms, and using Stirling's approximation for the binomial coefficients gives, from (A.2),

$$\lim_{N\to\infty} \left\{1 - \frac{1}{N} \log V_N(k_1)\right\} = 1 - E_1 = C_1 \qquad (A.8)$$

from (A.3), for $p_o < p_1 < \frac{1}{2}$,

$$\lim \sup_N \alpha_{opt}(N, p_o, p_1) \leqslant \lim_{N\to\infty} \frac{-\log P_{II}(N, p_o, p_1)}{N}$$

$$= -\Delta - \delta \log \frac{p_o}{q_o} \qquad (A.9)$$

and from (A.7), for $p_o < p_1 < p_{crit}$,

$$\lim \inf_N \alpha_{opt}(N, p_o, p_1) \geqslant \lim_{N\to\infty} \frac{-\log P_{III}(N, p_o, p_1)}{N}$$

$$= -\Delta - \delta \log \frac{p_o}{q_o} \qquad (A.10)$$

Together, Eqs. (A.9) and (A.10) prove the first part of the theorem, and cover the region in which the dashed-line and the dotted curves of Fig. 4 coincide.

Since the length represented by α_{opt} in this region is the difference between the curve and its tangent, it is second-order in δ or Δ. In fact, for small δ and Δ,

$$\alpha_{opt}(p_o, p_1) \approx \frac{\delta^2}{2pq} \log e \approx \frac{\Delta^2}{2pq\left(\log \frac{p}{q}\right)^2} \log e$$

For $p_1 > p_{crit}$, the largest term in the sum in Eq. (A.7) is not the last, but is that term for which $j^2/(N-j)^2$ is most nearly equal to p_o/q_o. This term is larger than $P_I(N, p_o, p_1)$ for large N, and the sum is bounded above by k_1, the number of terms, times the largest term. Taking the limit of (1/N) multiplied by the logarithm and using Stirling's approximation gives upper and lower bounds for $\alpha_{avg}(N, p_o, p_1)$ which coincide, giving for $p_{crit} < p_1 < \frac{1}{2}$,

$$\lim \inf_N \alpha_{opt}(N, p_o, p_1) \geqslant \lim_{N\to\infty} \alpha_{avg}(N, p_o, q_o)$$

$$= \lim_{N\to\infty} \frac{-\log P_{III}(N, p_o, p_1)}{N}$$

$$= C_{crit} + \alpha_{crit} - C \qquad (A.11)$$

This gives the remainder of the dotted curve in Fig. 4.

For p_1 less than p_{crit}, the probability of a detection error as computed above is essentially the probability of escaping from a sphere of radius $k_1 = Np_1$. For p_1 near 1/2, a different point of view is possible and leads to the two solid curves in Fig. 4.

The probability that transmission errors will cause one transmitted signal to be decoded as another is the probability that the noise will alter half or more of the positions in which they differ. If they differ in $Np_1 = k_1$ positions, this probability is just the upper half of the binomial, $P_I\left(Np_1, p_o, \frac{1}{2}\right)$ as given in Eq. (A.3). This is the probability of a particular transition; the total error probability is certainly less than this multiplied by the number of signal sequences. Gilbert[8] has shown that it is always possible to find $2^N/V_N(k_1 - 1)$ signal sequences each of which differs in at least k_1

positions from every other. For large N, by Eq. (A.8), this corresponds to a signaling rate of C_1 bits per symbol, or 2^{NC_1} signal points. Thus

$$P_e(N, p_o, p_1) \leqslant \frac{2^N}{V_N(k_1 - 1)} P_I\left(Np_1, p_o, \tfrac{1}{2}\right) \quad (A.12)$$

and asymptotically, from Eqs. (A.8) and (A.3),

$$\lim_{N\to\infty} \frac{-\log P_e(N,p_o,p_1)}{N} = -p_1 \left\{C - 0 + \left(\tfrac{1}{2} - p_o\right)\log \frac{p_o}{q_o}\right\}$$

$$= p_1 \cdot \frac{1}{2}\left\{\log \frac{1}{2p_o} + \log \frac{1}{2q_o}\right\}$$

$$= \frac{p_1}{2} \log \frac{1}{4pq} \quad (A.13)$$

and

$$\lim \inf_N \alpha_{opt}(N,p_o,p_1) \geqslant -C_1 + \frac{p_1}{2} \log \frac{1}{4pq} \quad (A.14)$$

This is the upper solid curve in Fig. 4. For an upper bound to α_{opt} we use a result of Plotkin[9] which shows that there are at most 2N signal points whose mutual minimum distance is as great as N/2. This means that the transmission rate for signal points at this distance is (1 + log N)/N and approaches zero for large N. This result sets a limit to the number of signal points at smaller distances as well. As Plotkin pointed out, if B(N,k) is the number of signal points at mutual distance ⩾k, then at least half of these agree in their first coordinate. Eliminating the n first coordinates gives

$$B(N-n, k) \geqslant 2^{-n} B(N, k) \quad (A.15)$$

Using Eqs. (A.14) and (A.15), let n = N - 2k. Then

$$B(N, k) \leqslant 2^{N-2k} B(2k, k) = 4k \cdot 2^{N-2k} \quad (A.16)$$

For a transmission rate C_1 this determines k:

$$C_1 = 1 - E_1 = \lim_{N\to\infty} \frac{\log B(N,k)}{N} \leqslant 1 - 2\frac{k}{N},$$

$$\text{or} \quad k \leqslant \frac{N}{2} E_1 \quad (A.17)$$

Now the error probability for such a set of signal points is certainly greater than the probability of a single transition, which is, in turn, at least as great as the upper half of the binomial

$$P_I\left(\frac{NE_1}{2}, p_o, \frac{1}{2}\right)$$

Thus

$$\lim \sup_N \alpha_{opt}(N,p_o,p_1) \leqslant \lim_{N\to\infty} \frac{-\log P_I\left(\frac{NE_1}{2}, p_o, \frac{1}{2}\right)}{N}$$

$$= \frac{E_1}{4} \log \frac{1}{4pq} \quad (A.18)$$

which gives the lower solid curve in Fig. 4. At $p_1 = 1/2$, Eqs. (A.18) and (A.14) give the same value, so that

$$\lim_{N\to\infty} \lim_{p_1\to 1/2} \alpha_{opt}(N, p_o, p_1) = \frac{1}{4} \log \frac{1}{4pq} \quad (A.19)$$

These results prove the remainder of the theorem. It should be noted that Eq. (A.19) does not imply that it is impossible to transmit any information with an error probability less than approximately $2^{-\frac{N}{4}\log\frac{1}{4pq}}$ for finite N. It is only impossible to do so while transmitting at a positive rate in the limit of large N. The transmission of one bit per block of N symbols can be accomplished by picking two signal sequences that differ in every position, with an error probability equal to $P_I(N, p_o, 1/2)$ for which, from Eq. (A.3),

$$\lim_{N\to\infty} \frac{-\log P_I\left(N, p_o, \frac{1}{2}\right)}{N} = \frac{1}{2} \log \frac{1}{4pq} \quad (A.20)$$

This error exponent is twice as great as that for the limit (as $p_1 \to 1/2$) of α_{opt} for positive transmission rates. Other points for the transmission of 2, 3, ... log N bits per block of N symbols fall between the value of Eq. (A.20) and that of Eq. (A.19).

2. Outline Proof of Theorems 4 and 5

A large part of the proof of Theorem 3 carries over directly for Theorems 4 and 5. Any upper bound on α_{opt} for the best possible code is automatically an upper bound for the more restricted class of check-symbol codes. Thus Eqs. (A.9) and (A.18), the tangent line and the lower solid curve in Fig. 4, still apply. To get the upper solid curve, Eq. (A.14), it is necessary to show that Gilbert's result, and thus Eq. (A.12), holds for the kind of code considered. This is obvious for pcs codes; Gilbert's proof requires only trivial modifications in this case. Since the pcs codes are a special case of check-symbol codes, the result follows for these as well. For sliding and convolutional pcs codes Gilbert's result is not obvious, although probably still true; that is why only the first part of Theorem 3 is extended to these cases.

The difficult point in Theorems 4 and 5 is the demonstration that the average of all possible codes, of each of the four types considered, is still given by Eqs. (A.10) and (A.11) and the dotted curve in Fig. 4. This requires a demonstration that the inequality of Eq. (A.6) still applies; that is, that the probability of a decoding error when j transmission errors have occurred is essentially the same, on the average, for the different types of check-symbol codes as for the average of all codes. The remainder of the derivation then follows as before.

This will be done for the pcs codes. When a noisy signal is received, the parity-check sums are recomputed at the receiver and added modulo two per position to the received check symbols, as in the Hamming code[7]. The resulting check-symbol pattern is the pattern caused by the transmission errors alone. The probability that this check-symbol pattern will be misinterpreted when j transmission errors have occurred is the probability that some other collection of j or fewer errors has the same check-symbol pattern. There are $V_N(j) - 1$ other patterns of j or fewer errors, and the probability that one of these has the same check-sum pattern is the probability that one of the $V_N(j) - 1$ differences has a check-sum pattern of all zeros. Now, if the check-sum matrix is filled in at random, any error pattern may produce any check-symbol pattern with equal probability. Therefore the probability of any one error pattern having a check-symbol pattern that vanishes is the reciprocal of the total number of possible check-symbol patterns. This number is $V_N(k_1)$, since $2^N/V_N(k_1)$ messages are being sent, and the total probability of a decoding error when j transmission errors have been made is less than this multiplied by the number of difference patterns:

$$P_{II}(N, j, k_1) \leqslant \frac{V_N(j) - 1}{V_N(k_1)} \leqslant \frac{V_N(j)}{V_N(k_1)} \leqslant \frac{\binom{N}{j}}{\binom{N}{k_1}} \qquad (A.21)$$

which is the inequality of (A.6) obtained by a different route.

The essential point in this argument is that every transmission error pattern, in the ensemble of possible codes, may cause every check-symbol pattern, with equal probability. Given this, the rest of the argument presented above follows. This is easy, but tedious, to show for sliding and convolutional pcs coding; the proofs will be omitted here.

References

1. C. E. Shannon, "A mathematical theory of communication," Bell System Tech. J. 27, 379-423, 623-656 (1948).

2. C. E. Shannon, "Communication in the presenc of noise," Proc. I.R.E. 37, 10-21 (1949).

3. M. J. E. Golay, "Note on the theoretical efficiency of information reception with PPM," Proc. I.R.E. 37, 1031 (1949).

4. R. M. Fano, "Communication in the presence of additive Gaussian noise," "Communication Theory," Willis Jackson Ed. (Butterworths, London, 1953) 169-182.

5. S. O. Rice, "Communication in the presence of noise -- probability of error of two encoding schemes," Bell System Tech. J. 29, 60-93 (1950).

6. A. Feinstein, "A new basic theorem of information theory," Trans. I.R.E. (PGIT) 4, 2-22 (1954).

7. R. W. Hamming, "Error detecting and error correcting codes," Bell System Tech. J. 29, 147-160 (1950).

8. E. N. Gilbert, "A comparison of signalling alphabets," Bell System Tech. J. 31, 504-522 (1952).

9. M. Plotkin, "Binary codes with specified minimum distance," Univ. of Penna., Moore School Research Division Report 51-20 (1951).

10. M. J. E. Golay, "Binary coding," Trans. I.R.E. (PGIT) 4, 23-28 (1954).

11. A. E. Laemmel, "Efficiency of noise-reducing codes," pp. 111-118 in "Communication Theory reference 4 above.

12. D. E. Muller, "Metric properties of Boolean algebra and their application to switching circuits," University of Illinois, Digital Computer Laboratory Report No. 46.

13. I. S. Reed, "A class of multiple error-correcting codes and the decoding scheme," Trans. I.R.E. (PGIT) 4, 38-49 (1954).

14. P. Elias, "Error-free coding," Trans. I.R.E. (PGIT) 4, 30-37 (1954).

15. W. Feller, "An Introduction to Probability Theory and Its Applications" (John Wiley and Sons, Inc., New York, 1950).

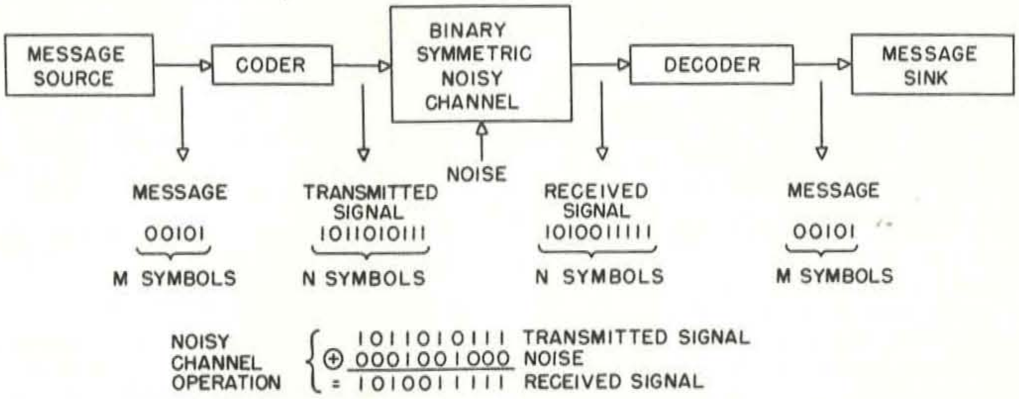
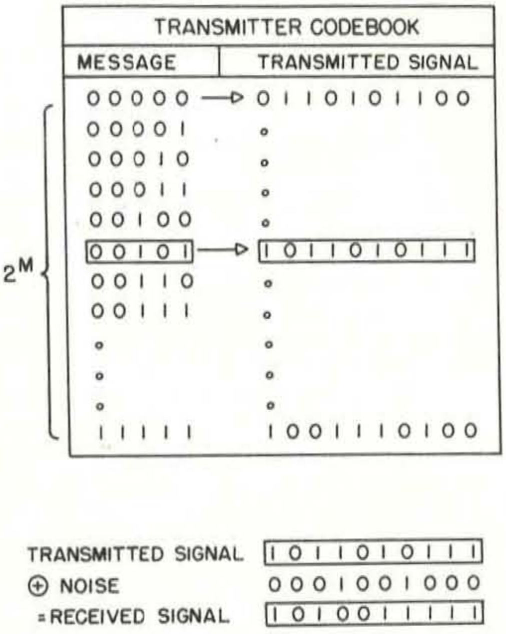


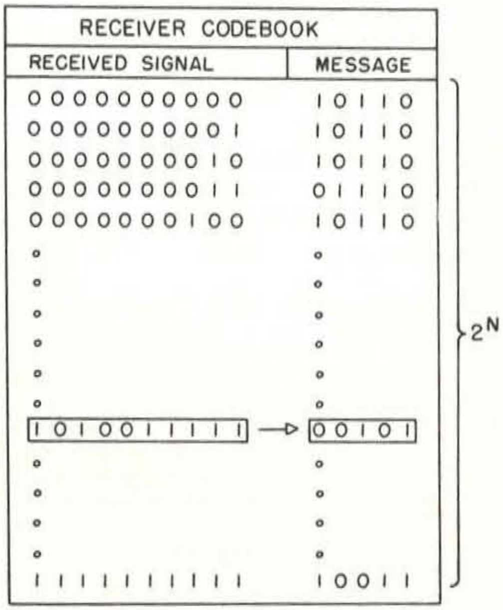
Fig. 1
Noisy communication system



MESSAGE	TRANSMITTED SIGNAL
00000	0110101100
00001	∘
00010	∘
00011	∘
00100	∘
00101	1011010111
00110	∘
00111	∘
∘	∘
∘	∘
∘	∘
11111	1001110100

TRANSMITTED SIGNAL	1011010111
⊕ NOISE	0001001000
= RECEIVED SIGNAL	1010011111

Fig. 2
Codebook coding



RECEIVED SIGNAL	MESSAGE
0000000000	10110
0000000001	10110
0000000010	10110
0000000011	01110
0000000100	10110
∘	∘
∘	∘
∘	∘
∘	∘
∘	∘
∘	∘
1010011111	00101
∘	∘
∘	∘
∘	∘
∘	∘
1111111111	10011

Fig. 3
Codebook decoding

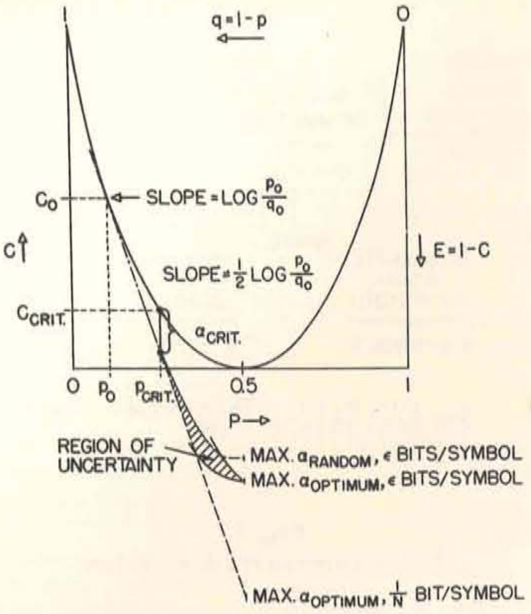


Fig. 4
The error exponent α optimum

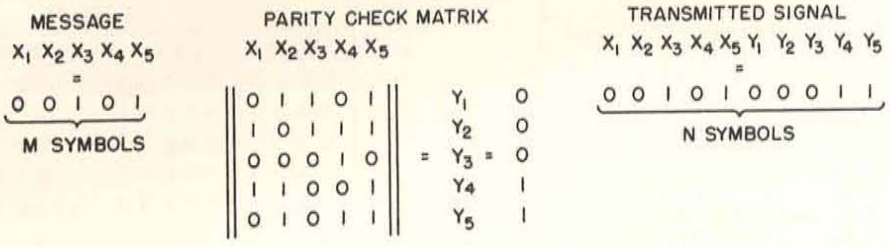


Fig. 5
Parity check coding

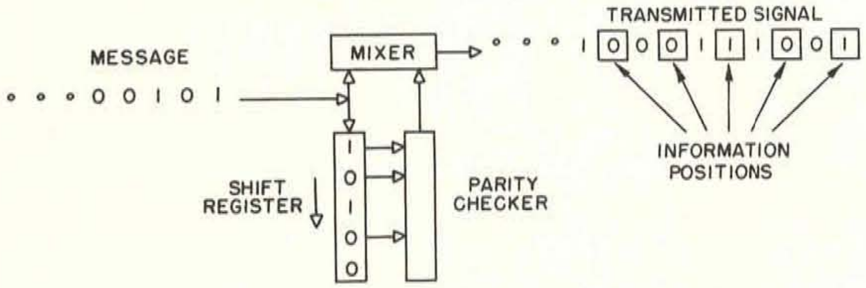


Fig. 6
Convolutional parity check coding