# Chapter 1

# INTRODUCTION

## 1.1 Coding for Digital Data Transmission

The need for efficient and reliable digital data communication systems has been rising rapidly in recent years. This need has been brought on by a variety of reasons, among them being the increase in automatic data processing equipment and the increased need for long range communication. Attempts to develop data systems through the use of conventional modulation and voice transmission techniques have generally resulted in systems with relatively low data rates and high-error probabilities.

A more fundamental approach to the problems of efficiency and reliability in communication systems is contained in the Noisy Channel Coding theorem developed by C. E. Shannon[15,4] in 1948. In order to understand the meaning of this theorem, consider Figure 1.1. The source produces binary digits, or binits, at some fixed time rate $R_t$. The encoder is a device that performs data processing, modulation, and anything else

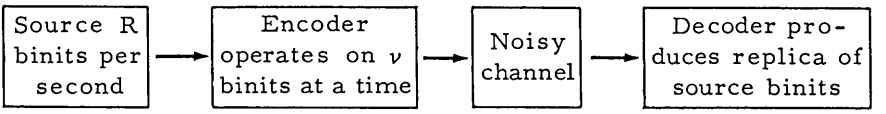


Figure 1.1. Block diagram of communication system.

that might be necessary to prepare the data for transmission over the channel. We shall assume, however, that the encoder separates the source sequence into blocks of $\nu$ binits and operates on only one block at a time. The encoder output is then transmitted over the channel and changed by some sort of random disturbance or noise. The decoder processes the channel output and produces a delayed replica of the source binits. The coding theorem states that for a large variety of channel models, encoders and decoders exist such that the probability of the decoder reproducing a source binit in error $P_e$ is bounded by

$$e^{-\nu[E_L(R_t)+0(\nu)]} \leq P_e \leq e^{-\nu E(R_t)}$$

The functions $E(R_t)$ and $E_L(R_t)$ depend upon the channel but

not upon $\nu$; they are positive when $R_t = 0$, and decrease with $R_t$ until they become 0 at some time rate $C_t$ known as the channel capacity. The exact nature of these functions and the particular class of channels for which this theorem has been proved need not concern us here. The important result is that the coding constraint length $\nu$ is a fundamental parameter of a communication system. If a channel is to be used efficiently, that is with $R_t$ close to $C_t$, then $\nu$ must be made correspondingly large to achieve a satisfactory error probability.

The obvious response of an engineer to such a theorem is: "Splendid, but how does one build encoders and decoders that behave in this way when $\nu$ is large?" It is rather sobering to observe that if an encoder stores a waveform or code word for each possible block of $\nu$ binits, then the storage requirement must be proportional to $2^{\nu}$, which is obviously impractical when $\nu$ is large. Fortunately, Elias[3] and Reiffen[13] have proved that for a wide variety of channel models, the results of the Noisy Channel Coding theorem can be achieved with little equipment complexity at the encoder by the use of parity-check coding. This will be described in more detail later.

Unfortunately, the problem of decoding simply but effectively when $\nu$ is large appears to be much more difficult than the problem of encoding. Enough approaches to this problem have been developed to assure one that the Coding theorem has engineering importance. On the other hand these approaches have not been carried far enough for the design of an efficient, reliable data communication system to become a matter of routine engineering.

This monograph contains a detailed study of one of the three or four most promising approaches to simple decoding for long constraint length codes. The purpose of publishing this work is primarily to show how such a coding and decoding scheme would work and where it might be useful. Also, naturally, it is hoped that this will stimulate further research on the subject. Further mathematical analysis will probably be fruitless, but there are many interesting modifications of the scheme that might be made and much experimental work that should be done.

In order to prove mathematically some results about low-density parity-check codes, we shall assume that the codes are to be used on a somewhat restricted and idealized class of channels. It is obvious that results using such channel models can be applied only to channels that closely approximate the model. However, when studying the probability of decoding error, we are interested primarily in the extremely atypical events that cause errors. It is not easy to find models that approximate both these atypical events and the typical events. Consequently the analysis of codes on idealized channels can provide only limited insight about real channels, and such insight should be used with caution.

The channel models to be considered here are called symmetric binary-input channels. By this we mean a time-discrete channel for which the input is a sequence of the binary digits 0 and 1 and the output is a corresponding sequence of letters from a discrete or continuous alphabet. The channel is memoryless in the sense that given the input at a given time, the output at the corresponding time is statistically independent of all other inputs and outputs. The symmetry requirement will be defined precisely in Chapter 3, but roughly it means that the outputs can be paired in such a way that the probability of one output given an input is the same as that of the other output of the pair given the other input. The binary symmetric channel, abbreviated BSC, is a member of this class of channels in which there are only two output symbols, one corresponding to each input. The BSC can be entirely specified by the probability of a crossover from one input to the other output.

If a symmetric binary-input channel were to be used without coding, a sequence of binary digits would be transmitted through the channel and the receiver would guess the transmitted symbols one at a time from the received symbols. If coding were to be used, however, the coder would first take sequences of binary digits carrying the information from the source and would map these sequences into longer redundant sequences, called code words, for transmission over the channel. We define the rate R of such codes to be the ratio of the length of the information sequence to the length of the code word sequence. If the code words are of length n, then there are $2^{nR}$ possible sequences from the source that are mapped into n-length code words. Thus only a fraction $2^{-n(1-R)}$ of the $2^n$ different n-length sequences can be used as code words.

At the receiver, the decoder, with its knowledge of which sequences are code words, can separate the transmitted n-length code word from the channel noise. Thus the code word is mapped back into the nR information digits. Many decoding schemes find the transmitted code word by first making a decision on each received digit and then using a knowledge of the code words to correct the errors. This intermediate decision, however, destroys a considerable amount of information about the transmitted message, as discussed in detail for several channels in Reference 1. The decoding scheme to be described here avoids this intermediate decision and operates directly with the a posteriori probabilities of the input symbols conditional on the corresponding received symbols.

The codes to be discussed here are special examples of parity-check codes.* The code words of a parity-check code are formed

* For a more detailed discussion of parity-check codes, see Peterson.[12]

by combining a block of binary-information digits with a block of check digits. Each check digit is the modulo 2 sum* of a prespecified set of information digits. These formation rules for the check digits can be represented conveniently by a parity-check matrix, as in Figure 1.2. This matrix represents a set of linear homogeneous modulo 2 equations called parity-check equations, and the set of code words is the set of solutions of these equations. We call the set of digits contained in a parity-check equation a parity-check set. For example, the first parity-check set in Figure 1.2 is the set of digits (1, 2, 3, 5).

The use of parity-check codes makes coding (as distinguished from decoding) relatively simple to implement. Also, as Elias[3] has shown, if a typical parity-check code of long block length is used on a BSC, and if the code rate is between critical rate and channel capacity, then the probability of decoding error will be almost as small as that for the best possible code of that rate and block length.

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 1 | 1 | 0 | 1 | 0 | 0 | $x_5 = x_1 + x_2 + x_3$ |
| $n(1 - R)$ | 1 | 1 | 0 | 1 | 0 | 1 | 0 | $x_6 = x_1 + x_2 + x_4$ |
| | 1 | 0 | 1 | 1 | 0 | 0 | 1 | $x_7 = x_1 + x_3 + x_4$ |

**Figure 1.2. Example of parity-check matrix.**

Unfortunately, the decoding of parity-check codes is not inherently simple to implement; thus we must look for special classes of parity-check codes, such as described in Section 1.2, for which reasonable decoding procedures exist.

## 1.2 Low-Density Parity-Check Codes

Low-density parity-check codes are codes specified by a matrix containing mostly 0's and relatively few 1's. In particular, an (n, j, k) low-density code is a code of block length n with a matrix like that of Figure 2.1, where each column contains a small fixed number j of 1's and each row contains a small fixed number k of 1's. Note that this type of matrix does not have the check digits appearing in diagonal form as do those in Figure 1.2. However, for coding purposes, the equations represented by these matrices can always be solved to give the check digits as explicit sums of information digits.

Low-density codes are not optimum in the somewhat artificial

---

*The modulo 2 sum is 1 if the ordinary sum is odd and 0 if the ordinary sum is even.

sense of minimizing the probability of decoding error for a given block length, and it can be shown that the maximum rate at which they can be used is bounded below channel capacity. However, the existence of a simple decoding scheme more than compensates for these disadvantages.

## 1.3 Summary of Results

An ensemble of (n, j, k) codes will be formed in Chapter 2, and this ensemble will be used to analyze the distance properties of (n, j, k) codes. The distance between two words in a code is simply the number of digits in which they differ. Clearly an important parameter in a code is the set of distances separating one code word from all the other code words. In a parity-check code, it can be shown that all code words have the same set of distances to the other code words.[12] Thus the distance properties for the ensemble can be summarized by the typical number of code words at each distance from the all-zero code word. It is found that the typical (n, j, k) code for $j \geq 3$ has a minimum distance that increases linearly with the block length for j and k constant. Figure 2.4 plots the ratio of minimum distance to block length for several values of j and k and compares the ratio with the same ratio for ordinary parity-check codes. The (n, j, k) codes with $j = 2$ exhibit markedly different behavior, and it is shown that the minimum distance of an (n, 2, k) code can increase at most logarithmically with the block length.

In Chapter 3, a general upper bound on the probability of decoding error for symmetric binary-input channels with maximum-likelihood decoding is derived for both individual codes and for arbitrary ensembles of codes. The bound is a function of the code only through its distance properties. The assumption of maximum likelihood decoding is made partly for analytic convenience and partly so as to be able to evaluate codes independently of their decoding algorithms. Any practical decoding algorithm, such as that described in Chapter 4, involves a trade-off between error probability and simplicity; the maximum-likelihood decoder minimizes the error probability but is totally impractical if the block length is large.

It is shown in Chapter 3 that if the distance properties of the code are exponentially related to the block length, and if the code rate is sufficiently low, then the bound to P(e) is an exponentially decreasing function of the block length. For the appropriate ensemble of codes, these bounds reduce to the usual random coding bounds.[3,4]

For the special case of the binary symmetric channel, a particularly simple bound to P(e) is found; this is used to show that over a range of channel crossover probabilities, a typical low-density code has the same error behavior as the optimum code

of a slightly higher rate. Figure 3.5 illustrates this loss of effective rate associated with low-density codes.

In Chapter 4, two decoding schemes are described. In the first, which is particularly simple, the decoder first makes a decision on each digit, then computes the parity checks and changes any digit that is contained in more than some fixed number of unsatisfied parity-check equations. The process is repeated, each time using the changed digits, until the sequence is decoded. The second decoding scheme is based on a procedure for computing the conditional probability that an input symbol is a 1; this is conditioned on all the received symbols that are in any of the parity-check sets containing the digit in question. Once again, the procedure is iterated until the sequence is decoded. The computation per digit per iteration in each scheme is independent of code length. The probabilistic, or second scheme, entails slightly more computation than the first scheme, but decodes with a lower error probability.

A mathematical analysis of the probability of decoding error using probabilistic decoding is difficult because of statistical dependencies. However, for a BSC with sufficiently small crossover probabilities and for codes with $j \geq 4$, a very weak upper bound to the probability of error is derived that decreases exponentially with a root of the code length. Figure 3.5 plots crossover probabilities for which the probability of decoding error is guaranteed to approach 0 with increasing code length. It is hypothesized that the probability of decoding error actually decreases exponentially with block length, while the number of iterations necessary to decode increases logarithmically.

Chapter 5 extends all the major results of Chapters 2, 3, and 4 to nonbinary low-density parity-check codes. Although the theory generalizes in a very natural way, the expressions for minimum distance, error probability, and probabilistic decoding performance error are sufficiently complicated that little insight is gained into the advantages or disadvantages of a multilevel system over a binary system. Some experimental work would be helpful here in evaluating these codes.

Some experimental results for binary low-density codes are presented in Chapter 6. An IBM 7090 computer was used to simulate both probabilistic decoding and the noise generated by several different types of channels. Due to limitation on computer time, the only situations investigated were those in which the channel was sufficiently noisy to yield a probability of decoding error greater than $10^{-4}$. The most spectacular data from these experiments are given in Figure 6.8, which emphasizes the advantages of a decoding scheme that operates from a likelihood receiver instead of a decision receiver.

## 1.4 Comparison with Other Schemes

Some other coding and decoding schemes that appear extremely promising for achieving low-error probabilities and high data rates at reasonable cost are the following: first, convolutional codes[3] with sequential decoding as developed by Wozencraft,[17] Fano,[5] and Reiffen;[14] second, convolutional codes with Massey's threshold decoding;[10] and third, the Bose-Chaudhuri codes[2] with the decoding schemes developed by Peterson[12] and Zierler and Gorenstein.[18]

It has been shown by Fano[5] that for arbitrary discrete memoryless channels, sequential decoding has a probability of decoding error that is upper bounded by a function of the form $e^{-\alpha n}$. Here n is the constraint length of the code and $\alpha$ is a function of both the channel and the code; $\alpha$ is positive for rates below channel capacity C. Fano also shows that for rates below a certain quantity called $R_{comp}$, where $R_{comp} < C$, the average amount of computation in decoding a digit is bounded by a quantity independent of constraint length.

An experimental sequential decoder has been built at Lincoln Laboratories, Lexington, Massachusetts.[11] By using this decoder in a system with a feedback link and an appropriately designed modulator and demodulator, reliable transmission has been achieved experimentally[9] over a telephone circuit at about 7500 bits per second rather than the 1200 or 2400 bits per second possible without coding.

The two principal weaknesses of sequential decoding are as follows: First, the amount of computation required per digit is a random variable, and this creates a waiting line problem at the decoder; second, if the decoder once makes an error, a large block of errors can be made before the decoder gets back on the proper track. If a feedback link is available, these problems are not serious, but considerably more study is required for cases in which no feedback exists.

Threshold decoding is the simplest scheme to implement that is discussed here; it involves only shift registers, a few binary adders, and a threshold device. It is most effective at relatively short constraint lengths, and has a somewhat higher error probability and less flexibility than sequential decoding.

The computation per digit associated with the Bose-Chaudhuri codes on the BSC increases roughly as the cube of the block length but does not fluctuate widely. The decoding scheme guarantees correction of all combinations of up to some fixed number of errors and corrects nothing beyond. For moderately long block lengths, this restriction in the decoding procedure causes a large increase in $P_e$. No way is known to make use of the a posteriori probabilities at the output of more general binary

input channels. This inability to make use of a posteriori probabilities appears to be a characteristic limitation of algebraic as opposed to probabilistic decoding techniques.

The computation per digit associated with low-density parity-check codes appears to increase at most logarithmically with block length and not to fluctuate widely with the noise. The probability of decoding error is unknown, but is believed to decrease exponentially with block length at a reasonable rate. The ability to decode the digits of a block in parallel makes it possible to handle higher data rates than is possible with the other schemes.

For many channels with memory, retaining the a posteriori probabilities from the channel makes it practically unnecessary to take account of the memory in any other way. For instance, on a fading channel when the fade persists for several baud lengths, the a posteriori probabilities will indicate the presence of a fade. If this channel were used as a BSC, however, it would be necessary for the decoder to account for the fact that bursts of errors are more probable than isolated errors. Then, using a posteriori probabilities gives low-density decoding and sequential decoding a great flexibility in handling channels with dependent noise. For channels in which the noise is rigidly constrained to occur in short, severe bursts, on the other hand, there is a particularly simple procedure for decoding the Bose-Chaudhuri codes.[12]

When transmitting over channels subject to long fades or long noise bursts, it is often impractical to correct errors in these noisy periods. In such cases it is advantageous to use a combination of error correction and error detection with feedback and retransmission.[16] All of the coding and decoding schemes being considered here fit naturally into such a system, but in cases where little or no error correction is attempted, low-density codes appear at a disadvantage.

In conclusion, all these schemes have their own advantages, and clearly no one scheme is optimum for all communication situations. It appears that enough coding and decoding alternatives now exist for serious consideration of the use of coding on particular channels.