Coding Theory

J.H. van Lint

1. Introduction

Coding theory, or more specifically, the theory of error-correcting codes is younger than the Foundation Mathematical Centre. We go back to the late 1940's. In those days computers were able to recognize bit errors (due to some technical failure) and if this happened, the process would be terminated. The idea is simple. Sequences of binary symbols (0 and 1) of a fixed length n are processed. Such sequences will be referred to as *words*. The only words that were allowed to be used were those with an even number of 1's. Clearly an error (which was luckily improbable) in a single bit would cause a violation of the parity rule and the computer stopped. In sufficiently long programs this would eventually happen. As a consequence R.W. Hamming of Bell Laboratories (USA) quite often found his computer not at work when he returned to the laboratories in the morning: an error had been *detected*. His irritation over the fact that an error could be detected but not corrected, led to his construction of the *Hamming code*, a so-called single-error-correcting code. The idea of the code can be easily understood from the following simple and elegant description, due to R.J. Mc Eliece. We assume that information is presented as a long string of 0's and 1's. These are to be communicated from a 'sender' to a 'receiver' over a medium that we shall call 'the channel'. This channel has the unpleasant property that, with a (luckily small) probability p, sometimes a 0 is changed into a 1 or a 1 into a 0 on the way to the receiver. We wish to significantly

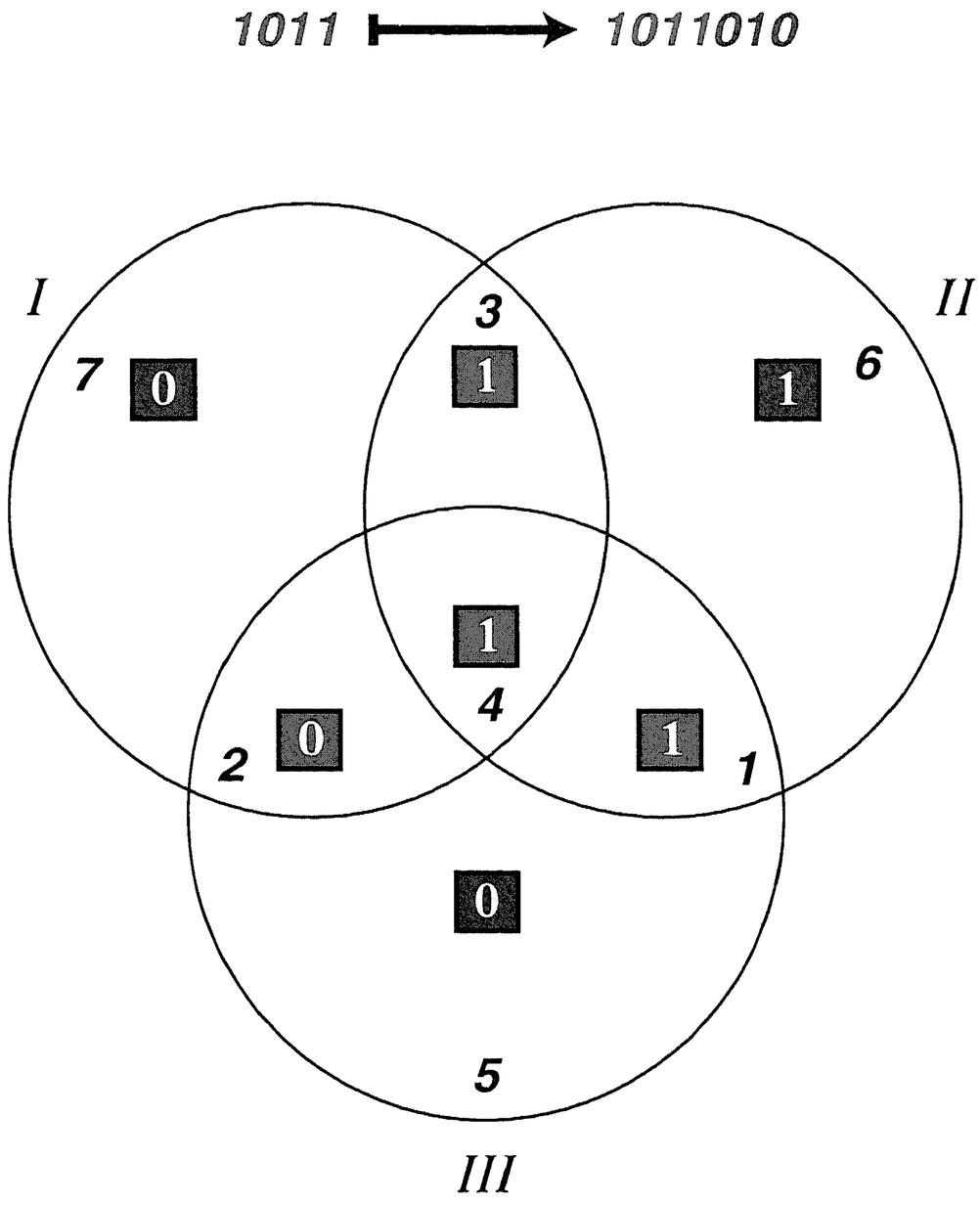


Figure 1. The Hamming code.

increase the probability of correct interpretation of the received sequence of 0's en 1's. For this we are willing to pay a toll (loss of time or energy or space, all depending on the practical application of this model). Here is what Hamming did. The sequence of 0's and 1's is divided into fourtuples, each of which is mapped into a seventuple. The seventuples are transmitted. The mapping is described in figure 1.

1.1. An example

Let (a_1, a_2, a_3, a_4) be a fourtuple. These bits are put into the positions 1,2,3,4, in the figure. From the positions 5,6,7 we find three so-called *redundant* bits by the following parity rule: each of the circles I,II,III should contain an even number of 1's. The reader will easily see that if one bit is received incorrectly, the receiver can see from figure 1 which circles violate the parity rule. This clearly identifies the position of the erroneous bit and *correction* can take place! To understand why this became an extremely exciting area in mathematics, we must analyse the '*code*' described above. This will allow us to quote the theorem that started coding theory. It is known as Shannon's Theorem, put forward in his monumental paper in 1948 (see [1]).

In the example above, each received sequence consists of four bits of *information* and three redundant bits. We say that the *information rate* R equals 4/7. As an exercise, the reader can check that if the probability p that a bit is transmitted erroneously by the channel is say 0.001, then the probability that a fourtuple is misinterpreted by the receiver is roughly 2.10^{-5}. Clearly not using coding would imply an error probability of 4.10^{-3}. This is an impressive improvement in error probability and the toll is a decrease of an information rate from 1 to 4/7. It may be of interest to the reader to know that, on the most impressive application of coding theory, to wit the *compact disc*, the information rate is 3/4, i.e., one fourth of the disc does not contain music but redundancy added by coding theorists, responsible for the superb quality of the music!

The channel that we described above is known as the *binary symmetric channel*. The model assumes that a bit-error is a random event with a given probability p. For such a channel, we define the *capacity* C as $C = 1 + p \log p + (1-p)\log(1-p)$. (Logarithms to base 2.) In our example, we have $C \approx 0.99$.

Shannon's theorem states that for a binary symmetric channel with capacity C and for any $\varepsilon > 0$ and any $R < C$, there exists a code with rate at least R and error probability (after decoding) less than ε. This sounds unbelievable. What one should realize is that the codes of this theorem are *extremely long*.

1.2. Applications

Where has coding theory gone? We mention some of the important areas of application (see also figure 2). As mentioned in the introduction, computing is an important area of application. Since most modern communication is digital, error-correction plays a role there. Spectacular applications were the photographs taken by several satellite missions (Mars, Saturn, Jupiter). Without coding theory there would not have been pictures at all. In recent times the CD is the most notable application. In these examples, the channels are quite different. For telephone, it is light in optical fibre, for satellites radio communication (where the source of errors is thermal noise in the amplifier at the receiver), for CD the errors are caused by dust, air bubbles in the disc, scratches, finger prints, etc. Practical requirements can differ considerably: processing the signals from Mariner Mars took one day; the CD-player has a delay (for decoding) of a fraction af a second. Obviously, a lot of energy has gone into finding good (i.e., fast) decoding algorithms.

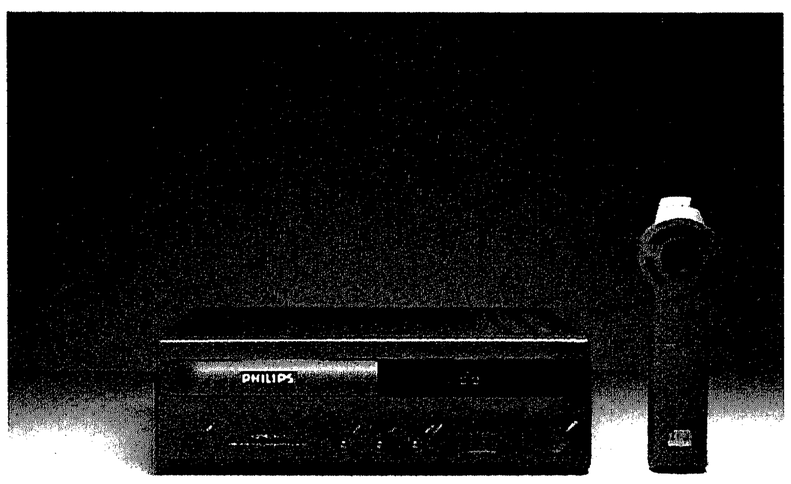
1.3. Parameters

To understand some of the mathematical developments, we need some parameters. We use n for the length of the code. The alphabet is not necessarily $\{0, 1\}$. We use q for the size of the alphabet. In algebraic coding theory, q is a prime power and the alphabet is a finite field. (For CD we have $q = 2^8$.) Up to now, we have not mentioned the most important parameter of a code, to wit its *distance*. The distance of two words is the number of places where they differ (e.g., 101011 and 100010 have distance 2, since they differ in positions three and six). The *minimum distance* d of a code C is the minimum value of the distance between two distinct codewords. Obviously for a code with minimum distance $d = 2e + 1$, it is theoretically possible to correct up to e errors. The most important problem in combinatorial Coding Theory is to establish bounds for the number $A_q(n, d)$, the *maximal*

number of codewords in a code C of length n over an alphabet of size q, and with minimum distance d. For example $A_2(7, 3) = 16$ and the Hamming code described above realizes this bound. Clearly, for practical applications one has tried to find relatively good codes of moderate length, i.e., codes with given n, q, and d, for which $|C|$ is close to $A_q(n, d)$. Quite often, mathematical results are not appealing to engineers because the construction of the code gives no hint as to how the receiver can decode (quickly).

1.4. Covering codes

We now turn to so-called *covering codes*. This area of research is very important for error-correcting codes themselves but is also, in some sense, complementary. Some of the applications concern *data-compression* (important for high definition television). Here the problem is that we have



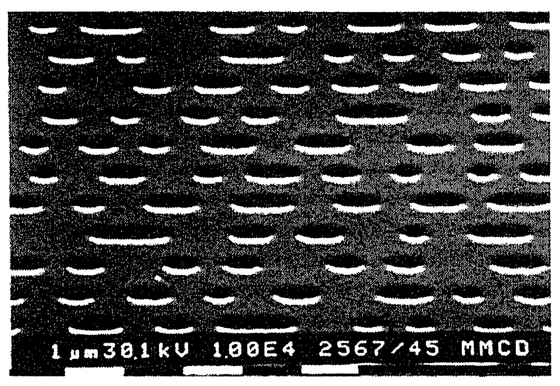


Figure 2. Coding theory is applied in several areas. Spectacular examples include: Voyager pictures of Saturn (above) and CD-players (below, high density pit structure shown on the right; courtesy Philips NV Eindhoven).

too much information to transmit and that we are willing to accept small deviations in order to save time or space. We illustrate this using our earlier example of the Hamming code. We first introduce some geometric terminology. If $\mathbf{c}$ is a codeword in a code C, then the *ball* of radius r around $\mathbf{c}$ consists of all words $\mathbf{x}$, such that $d(\mathbf{x}, \mathbf{c}) \leq r$. Note that if C has minimum distance $d = 2e + 1$, then the balls of radius e around the codewords are disjoint. There can be many words $\mathbf{x}$ that are not in any of these balls. In practice of error-correction this is important. A received word that is within a ball of radius e around a codeword $\mathbf{c}$ is decoded as $\mathbf{c}$. If $\mathbf{x}$ is received and $\mathbf{x}$ is not in any of these balls, the receiver knows that too many errors have occurred for him to handle the situation.

Back to data compression. We have to send information consisting of seven bits. If the receiver makes one error, this is close enough for him to interpret our meaning. (In TV this is achieved by the redundancy of the overall picture.) For each seventuple we consider the ball in the Hamming code to which it belongs, determine the centre and transmit only the bits a_1, a_2, a_3, a_4 corresponding to that centre. The receiver is never more than one bit off (for the seventuple) and we have only 4/7 of the amount of information to be transmitted, a considerable gain. This leads to the definition of the *covering radius* of a code C (a subset of all words of length n over an alphabet of size q). The covering radius is the smallest number $\boldsymbol{\partial}$ such that the balls of radius $\boldsymbol{\partial}$ around the codewords of C cover all words of length n. For the Hamming code we have $\boldsymbol{\partial} = 1$. This code is called *perfect* because the balls of radius 1 around codewords are not only disjoint but they also cover the space. In practice, perfect codes are far from perfect; the occurrence of too many errors should in general not escape the receiver. There has been far less research on covering codes than on error-correcting codes.

1.5. Coding theory and algebraic geometry

For a long time, coding theory was an area for electrical engineers. Many of their interesting achievements later turned out to be equivalent to mathematical methods and results that had been known for a long time. It was not until the 1970's that mathematicians became interested in coding theory (influenced partly by activities in The Netherlands). Especially the relations between coding theory and design theory (an area with its origin in statistics and quality control) led to a surge of interest. We will go into this below. Algebraic methods became increasingly important (especially through the work of Ph. Delsarte). A peculiar connection to simple groups pulled in another area, so that by the 1980's coding theory was respectable for group theorists, algebraists and of course combinatorialists. And then algebraic geometry appeared on the scene of coding theory, 'bien étonné de

se trouver ensemble'! It is extremely difficult to explain the connection but let's try.

Everyone is familiar with the fact that the rational numbers are a subfield of the reals. Everyone is familiar with the curve S known as the unit circle: $\{S = (x, y) : x^2 + y^2 = 1\}$. The point $(3/5, 4/5)$ on S has rational coordinates but in general the coordinates on S will be irrational. We are interested in codes over an alphabet F_q (the finite field with q elements). This field is a subfield of its algebraic closure F (which is an infinite field). In algebraic geometry one studies curves (defined using coordinates in F), given by an algebraic equation (like the circle S above). One of the important problems is to determine the 'rational' points on S. Here, rational means that the coordinates are in the subfield F_q.

We now describe the link to coding theory. Let S be an algebraic curve over F with n rational points $P_1, P_2, \ldots, P_n$. Consider a suitably chosen set $\mathcal{F}$ of rational functions defined on S. The code C is defined as the collection of words $(f(P_1), f(P_2), \ldots, f(P_n))$ of length n obtained by letting f run through $\mathcal{F}$. If we know enough about the curve S, it is possible to make (interesting) assertions about the code C, i.e., about its minimum distance. Using some deep results from algebraic geometry, M.A. Tsfasman, S.G. Vlădut, and Th. Zink in 1982 proved the existence of codes that are far better than anything that was believed possible until then. Clearly a sensational development. This result has led to quite a lot of research in which The Netherlands has made significant contributions. Three directions can be mentioned. First, studying special classes of curves to see if reasonably good codes can be found. Second, finding decoding algorithms, preferably good enough to get engineers interested. (Notice that the problem has reversed.) Recently, there has been progress in translating the results from algebraic geometry into terminology that avoids the deep mathematics but produces nearly the same results [2]. This is an exciting area that has just been opened.

2. Coding theory at CWI

Research on coding theory at CWI started in 1972. This marked the beginning of the strong collaboration with the Discrete Mathematics group at the Eindhoven University of Technology (TUE, which has lasted). Important contributions from the early years are several results on bounds on codes due to M.R. Best and A.E. Brouwer. In 1974 CWI organized the Advanced Study Institute on Combinatorics at Nijenrode Castle. It is still considered one of the major events in this area of the past 25 years! At the meeting Ph. Delsarte (invited speaker) presented his theory of the association schemes of coding theory. It has had a very strong influence on the research at both CWI and TUE. One of the open problems mentioned in the lecture of J.H. van Lint on perfect codes was solved by M.R. Best in his Ph.D. thesis

(1982). In 1975 a course on coding theory was held at the Mathematical Centre (CWI's name before 1983). This led to MC Syllabus 31 which was the basis for a book on Coding Theory [3].

3. PH.D. PROJECTS (NWO/SMC) PERFORMED AT TUE

One part of the Ph.D. thesis of H.J. Tiersma (1989) was concerned with codes from algebraic geometry, namely those based on Hermitean curves. The main part of the thesis concerns constructions and bounds for codes for channels that differ from the binary symmetric channel described above. We do not go into details but the idea is that more than one user uses the channel and information goes both ways, or it is added, etc.

G.J.M. van Wee (Ph.D. thesis 1991) also dedicated some of his time to algebraic geometry codes. The question that is answered is 'Which linear codes are algebraic-geometric?' (joint work with R. Pellikaan and B.Z. Shen). It turns out that in the class of codes known as Hamming codes only those with at most two redundant symbols can be constructed with an algebraic curve, with the exception of our example of figure 1.

The main part of the thesis concerns covering codes. One result deserves special mention. It is an elementary but ingenious counting argument that produces a lower bound on the number of words in a code of given length and covering radius. It is now known as the Van Wee bound. For this Ph.D. thesis Dr. van Wee was awarded the prestigious 'Dissertationspreis' of the Gesellschaft für Mathematik, Ökonomie und Operations Research in 1992.

In the meantime the algebraic-geometry code research group at the TUE, supervised by R. Pellikaan, had obtained international recognition. This led to several visits from researchers from abroad. An important further step in this process was the Ph.D. thesis by I.M. Duursma (1993) on Decoding Codes from Algebraic Curves. Mathematically speaking, the problem is solved but for practical use it is essential that far more efficient methods are found. This research (often jointly done with visitors) has led to sufficient

insight in these codes to be able to describe them in a more elementary way, thus opening the door to practical use. The decoding methods also led to new decoding techniques for cyclic codes. This extremely successful SMC-project led to 12 publications and 25 lectures abroad by Dr. Duursma!

In 1994 two SMC-projects resulted in a Ph.D. thesis. Feng-Wen Sun constructed decoding techniques and a modulation scheme for band-limited communications. The channels concerned differ considerably from those above. Either the signal is continuous and noise is Gaussian noise and not discrete, or the errors are not discrete but weighted in some way. Despite this different approach to signalling there is strong interplay with traditional coding theory in the thesis.

R. Struik extended the work of Van Wee on covering codes. In fact, he gave an improvement of the Van Wee bound. This led to several new

records. Furthermore the thesis analyses the codes with covering radius 2 or 3 and presents several new constructions for covering codes.

Both for the area of covering codes and for codes from algebraic geometry it is clear that the results mentioned above have made it possible to formulate many new interesting projects. It is therefore quite desirable that this sequence of SMC-projects is continued in the future.

REFERENCES

1. C.E. SHANNON (1948). A mathematical theory of communication. *Bell Syst. Tech. J. 27*, 379-423 and 623-656.
2. T. HØHOLDT, J.H. VAN LINT, R. PELLIKAAN (1996). Algebraic geometric codes. R.A. BRUALDI, W.C. HUFFMAN, V. PLESS (eds.). *Chapter of Handbook of Coding Theory*, Elsevier Science Publishers, Amsterdam, to appear.
3. J.H. VAN LINT (1982). *Introduction to Coding Theory*, Springer-Verlag.