# Chapter 2

# Coding theory

## 1. Motivation: transmission of messages

Let us consider the following situation. Person $A$ is in a space-craft somewhere in space. They navigate the space-craft according to the instructions that are received from person $B$ who is on Earth. For simplicity let us assume that there are four possible instructions: go left, go right, go up and go down. These instructions are transmitted as a binary radio-signal; in other words $B$ can transmit either of two types of signals, which we denote by 0 and 1.

So we have to encode four 'messages' into 'words' of 0s and 1s. The first, most obvious way is to do something like this:

$$\text{left}=00,\ \text{right}=01,\ \text{up}=10,\ \text{down}=11.$$

Once transmitted, the signal is affected by various factors, such as other radio-signals, cosmic rays etc. Thus there is a (small) chance that an emitted 0 will be received as 1, or that an emitted 1 will be received as 0. We assume three things about the possibility that an error occurs:

(BSC1) The probability that a 0 is turned into a 1 is the same as the probability that a 1 is turned into a 0.

(BSC2) This probability $p$ is the same for each digit, and is less than 0.5.

(BSC3) An error occurring in one digit does not affect the probability that an error occurs in another digit.

Any communication channel satisfying BSC1–3 is called a *Binary Symmetric Channel* (BSC). In this course we will always assume that we have a BSC.

So assume that $B$ sends the signal 00 (left), but the first 0 is changed into 1, so that $A$ receives 10. Clearly $A$ has no indication that an error has occurred, as 10 is also a valid instruction, and so $A$ will go up, rather than left.

Consider another way of encoding our instructions:

$$\text{left}=000,\ \text{right}=011,\ \text{up}=101,\ \text{down}=110.$$

Suppose that 000 (left) is sent, and a single error occurs, say changing the first 0 into 1. This time $A$ will be able to detect it, as 100 is not a valid message. However, $A$ will not know what was the original message, even if $A$ knows that only one error has occurred. For 100 might have been obtained from 110, as well as from 000, by changing one symbol.

Consider a further example:

$$\text{left}=00000,\ \text{right}=01101,\ \text{up}=10110,\ \text{down}=11011.$$

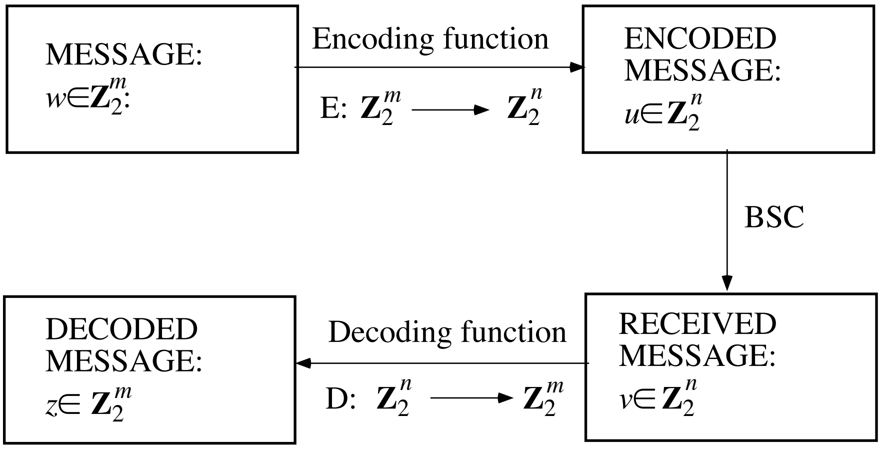

Figure 2.1: Transmission of messages

This time messages are sufficiently different so as to allow detection *and* correction of a single error. Thus, if $A$ receives the message 10000, and if it is assumed that only one error has occurred, than $A$ will know that the original message was 00000.

So, in essence, we are considering a scheme shown in Figure 2.1.

**Definition 1.1.** An *$n$-ary code* over $\mathbb{Z}_2$ is a subset $C \subseteq \mathbb{Z}_2^n$. The elements of $C$ are called *code-words*. Given a code $C$, an *encoding function* is any bijection $E : \mathbb{Z}_2^m \longrightarrow C$. A *decoding function* is any function $D : \mathbb{Z}_2^n \longrightarrow \mathbb{Z}_2^m$ such that for $u \in \mathbb{Z}_2^m$ we have $uED = u$.

In the ideal situation we would like the following to happen: we take an arbitrary word $w \in \mathbb{Z}_2^m$, encode it, transmit it, then decode it, and we obtain the same word $w$. This is clearly impossible, since we have no control over what will happen with $w$ in the channel. So we want to ensure that we have a high chance of decoding the message correctly. In other words, we want to be sure that we will decode the received word correctly, under the assumption that only a few errors have occurred.

## 2. Hamming distance

Since we will be dealing with elements of $\mathbb{Z}_2^n$ throughout this chapter, let us recall that these elements are $n$-tuples of 0s and 1s. Usually, instead of $(x_1, x_2, \ldots, x_n)$ we shall simply write $x_1 x_2 \ldots x_n$. We shall also frequently refer to these elements as *words*, rather than *vectors*. We also recall the addition and multiplication in $\mathbb{Z}_2$: $0+0 = 1+1 = 0$, $0+1 = 1+0 = 1$, $0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0$, $1 \cdot 1 = 1$. Note that for every $x$ we have $x = -x$, as $x + x = 0$ over $\mathbb{Z}_2$.

The Hamming distance provides us with a means of measuring the difference between any two words from $\mathbb{Z}_2^n$.

**Definition 2.1.** Let $x = x_1 x_2 \ldots x_n$ and $y = y_1 y_2 \ldots y_n$ be two words from $\mathbb{Z}_2^n$. The *Hamming distance* $d(x, y)$ between $x$ and $y$ is the number of places in which they differ:

$$d(x, y) = |\{i \,:\, 1 \leq i \leq n,\ x_i \neq y_i\}|.$$

Closely related to this is the notion of weight.

**Definition 2.2.** The *weight* of a word $x = x_1x_2\dots x_n \in \mathbb{Z}_2^n$, denoted by $wt(x)$, is the number of 1's in $x$:

$$wt(x) = |\{i \,:\, 1 \le i \le n,\ x_i = 1\}|.$$

The connection between distance and weight is as follows.

**Theorem 2.3.** *For any two $x, y \in \mathbb{Z}_2^n$ we have*

*(i)* $d(x, y) = wt(x - y)$*;*

*(ii)* $wt(x) = d(x, 0)$*,*

*where* $0$ *denotes the zero-vector in* $\mathbb{Z}_2^n$*.*

**Proof.** We have

$$d(x, y) = |\{i \,:\, x_i \ne y_i\}| = |\{i \,:\, x_i - y_i \ne 0\}| = |\{i \,:\, x_i - y_i = 1\}| = wt(x - y).$$

The proof of (ii) is similar. ■

Next we prove that the Hamming distance has the usual properties of a distance function:

**Theorem 2.4.** *The set $\mathbb{Z}_2^n$ with the Hamming distance $d$ is a metric space. In other words, $d$ has the following properties:*

*(i)* $d(x, y) \ge 0$*;*

*(ii)* $d(x, y) = 0 \iff x = y$*;*

*(iii)* $d(x, y) = d(y, x)$*;*

*(iv)* $d(x, z) \le d(x, y) + d(y, z)$ *(the triangle inequality);*

*for all $x, y, z \in \mathbb{Z}_2^n$.*

**Proof.** Properties (i), (ii) and (iii) are obvious, and we leave the proofs as an exercise. For (iv) note that

$$\{i \,:\, x_i \ne z_i\} \subseteq \{i \,:\, x_i \ne y_i \text{ or } y_i \ne z_i\} = \{i \,:\, x_i \ne y_i\} \cup \{i \,:\, y_i \ne z_i\},$$

so that

$$d(x, z) = |\{i \,:\, x_i \ne z_i\}| \le |\{i \,:\, x_i \ne y_i\}| + |\{i \,:\, y_i \ne z_i\}| = d(x, y) + d(y, z),$$

as required. ■

**Definition 2.5.** The *minimum distance* of a code $C$ is the minimum distance between any two codewords of $C$.

Now let us again consider a typical transmission process, where we have a code $C \subseteq \mathbb{Z}_2^n$, and where a word $u \in C$ has been transmitted through the channel. We know that errors may occur, and so, in general, the received word $v$ will be distinct from $u$. If we let $x = v - u (= v + u)$ we say that $x$ is the *error* of transmission. From $x = v - u$ we clearly have $v = u + x$, and so we say that the channel has added the error $x$ to the transmitted word $u$.

Let $X \subseteq \mathbb{Z}_2^n$ be arbitrary. We think of $X$ as the collection of errors which are more likely to occur than the others. We say that we can *detect* errors from $X$ if for every code-word $u \in C$ we have $u + x \notin C$. Similarly, we say that we can *correct* errors from $X$ if for all $u, u_1 \in C$ and all $x, x_1 \in X$ the equality $u + x = u_1 + x_1$ implies $u = u_1$ (and $x = x_1$). This means that no received word could have been produced by adding errors in $X$ to two different code-words. There is a strong connection between the Hamming metric and the error-detecting and error-correcting capabilities of a code.

**Theorem 2.6.** *Let $C \subseteq \mathbb{Z}_2^n$ be a code, and let $k \geq 1$.*

*(i) We can detect every error of weight at most $k$ if and only if $C$ has minimum distance at least $k+1$.*

*(ii) We can correct every error of weight at most $k$ if and only if $C$ has minimum distance at least $2k+1$.*

**Proof.** (i) ($\Rightarrow$) Suppose that we can detect every error of weight at most $k$. Let $u, v \in C$, $u \neq v$. Note that $v = u + (v - u) \in C$, so that we cannot detect the error $v - u$. Hence $k < wt(v - u) = d(u, v)$, and the minimum distance is at least $k + 1$.

($\Leftarrow$) Suppose that $C$ has minimum distance $k+1$. Let $u \in C$ and let $wt(x) \leq k$. Consider the word $u+x$. We have $d(u, u+x) = wt(u+x-u) = wt(x) \leq k$, so that $u + x \notin C$, and we can detect $x$. Thus we can detect every error of weight at most $k$.

(ii) ($\Rightarrow$) Let us assume that we can correct every error of weight at most $k$, but that there are two code-words $u, v \in C$ such that $d(u, v) \leq 2k$. If we let $I = \{i : u_i \neq v_i\}$, then $|I| \leq 2k$, and hence we can write $I$ as the union of two disjoint subsets $I_1$ and $I_2$ of size at most $k$:

$$I = I_1 \cup I_2,\ |I_1| \leq k,\ |I_2| \leq k,\ I_1 \cap I_2 = \emptyset.$$

Define two words $x = x_1x_2 \ldots x_n$ and $y = y_1y_2 \ldots y_n$ in $\mathbb{Z}_2^n$ as follows:

$$\begin{aligned} x_i = 1 &\Leftrightarrow i \in I_1, \\ y_i = 1 &\Leftrightarrow i \in I_2. \end{aligned}$$

Clearly we have $wt(x) \leq k$, $wt(y) \leq k$ and $u + x + y = v$. The last equality can be written as $u + x = v + y$, and, because of the error correcting capability of $C$, we conclude that $u = v$, a contradiction. So the minimum distance of the code is at least $2k + 1$.

($\Leftarrow$) Suppose that $C$ has minimum distance at least $2k + 1$. Let $u, v \in C$ and $x, y \in \mathbb{Z}_2^n$ be such that $wt(x) \leq k$, $wt(y) \leq k$ and $u + x = v + y$. Then we have

$$d(u, v) = wt(u - v) = wt(y - x) = wt(x + y) \leq wt(x) + wt(y) \leq 2k.$$

Since the minimum distance is at least $2k + 1$, we conclude that $u = v$, meaning that we can correct all errors of weight at most $k$. ■

**Example 2.7.** Consider the code $C = \{00001, 01010, 10100, 11111\} \subseteq \mathbb{Z}_2^5$. The distances between elements of $C$ are respectively 3, 3, 4, 4, 3, 3, and so we can detect errors of weight at most 2, and correct errors of weight 1.

## 3. Linear codes

We have seen that a code is simply a subset $C \subseteq \mathbb{Z}_2^n$, that an encoding function is a bijection $E : \mathbb{Z}_2^m \longrightarrow C$ and a decoding function is a mapping $D : \mathbb{Z}_2^n \longrightarrow \mathbb{Z}_2^m$ such that $uED = u$ for all $u \in \mathbb{Z}_2^n$. The problem with these general codes is that the encoding and decoding functions are not convenient for computing. For instance, for the encoding function one has to store a table containing all the elements of $\mathbb{Z}_2^m$ and the corresponding elements of $C$, and to look in this table whenever sending a message. This problem can be overcome by giving codes an algebraic structure, most often that of a vector space.

**Definition 3.1.** A *linear code* is any subspace of the vector space $\mathbb{Z}_2^n$.

The first advantage of having a linear code, as opposed to an arbitrary code, is that it is easier to analyse its error-detecting and error-correcting capabilities.

**Theorem 3.2.** *Let $C \subseteq \mathbb{Z}_2^n$ be a linear code. Then the minimum distance is equal to the minimal weight of a non-zero vector in $C$. In particular, we can detect (respectively, correct) every error of weight at most $k$ if and only if this minimal weight is at least $k+1$ (respectively, $2k+1$).*

**Proof.** Let $M$ be the minimum distance, with $d(x,y)=M$, and let $N$ be the minimal weight of a non-zero vector, with $wt(z)=N$. Since $C$ is a subspace of $\mathbb{Z}_2^n$ we must have $x-y\in C$ and also $0\in C$. But then we have

$$M=d(x,y)=wt(x-y)\geq N=wt(z)=d(z,0)\geq M,$$

and so $M=N$ as required. ■

Assume that we want to encode elements of $\mathbb{Z}_2^m$ by means of a linear code $C\subseteq\mathbb{Z}_2^n$. Then $C$ is a subspace of $\mathbb{Z}_2^n$ with $2^m$ elements, and so $\dim(C)=m$. Hence we can define $C$ by listing a basis for $C$, which is a set of $m$ linearly independent vectors from $C$:

$$a_i=a_{i1}a_{i2}\dots a_{in},\ 1\leq i\leq m,$$

If we take these vectors for the rows of a matrix $G$, we obtain what is called a *generator matrix* for $C$:

$$G=\begin{bmatrix} a_1\\ a_2\\ \vdots\\ a_m\end{bmatrix}=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22} & \dots & a_{2n}\\ \vdots & \vdots & & \vdots\\ a_{m1} & a_{m2} & \dots & a_{mn}\end{bmatrix}.$$

It is worth remarking that $G$ is not unique, as $C$ has several bases.

**Example 3.3.** Let $C:=\{000,110,101,011\}$. Then

$$G_1=\begin{bmatrix} 1&1&0\\ 0&1&1\end{bmatrix}$$

is a generator matrix for $C$, but so is

$$G_2=\begin{bmatrix} 1&0&1\\ 1&1&0\end{bmatrix}.$$

The generator matrix of a code can be used to define an easy encoding function.

**Theorem 3.4.** *Let $C\subseteq\mathbb{Z}_2^n$ be a linear code of dimension $m$, and let $G$ be a generator matrix for $C$. Then the function $E:\mathbb{Z}_2^m\longrightarrow\mathbb{Z}_2^n$ defined by $E(x)=xG$, is an encoding function.*

**Proof.** We have to prove that $E$ is a bijection. First note that for $x=x_1x_2\dots x_m\in\mathbb{Z}_2^m$ we have

$$E(x)=xG=[x_1\,x_2\dots x_m]\begin{bmatrix} a_1\\ a_2\\ \vdots\\ a_m\end{bmatrix}=x_1a_1+x_2a_2+\dots+x_ma_m\in C,$$

and so $E$ maps $\mathbb{Z}_2^m$ onto $C$, since $\{a_1,\dots,a_m\}$ is a basis for $C$. Also,

$$E(x)=E(y)\Rightarrow x_1a_1+\dots+x_ma_m=y_1a_1+\dots+y_ma_m\Rightarrow x_i=y_i\ (1\leq i\leq m),$$

since the vectors $a_1,\dots,a_m$ are linearly independent. Therefore $E$ is indeed a bijection. ■

**Example 3.5.** Let us consider the encoding function $E : \mathbb{Z}_2^3 \longrightarrow \mathbb{Z}_2^6$ given by the generator matrix $G = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$. Then we have $[1\,0\,1]G = [0\,0\,0\,1\,0\,1]$ and so the word 101 is encoded as 000101. In this way one can calculate all the code-words, and obtain $C = \{000000, 101100, 011011, 101001, 110111, 000101, 110010, 011110\}$. The weights of the code-words are respectively 0, 3, 4, 3, 5, 2, 3, 4. So we can detect single errors, but cannot correct them.

Another way to define an $m$-dimensional linear code in $\mathbb{Z}_2^n$ is to give it as the null-space of an $(n-m) \times n$ matrix with linearly independent rows. This matrix is called the *parity check matrix.*

**Example 3.6.** Consider the matrix

$$H = \left[\begin{array}{ccc} 1 & 1 & 1 \end{array}\right].$$

A vector $v = v_1v_2v_3 \in \mathbb{Z}_2^3$ is in the null-space of $H$ if and only if $v_1+v_2+v_3 = 0$. Thus the null-space of $H$ is $\{000, 110, 101, 011\}$, which is exactly the code of Example 3.3.

The question then arises about the connection between the generator matrix and the parity check matrix for the same code.

Let the code $C$ be given by the generator matrix $G = [a_{ij}]_{m\times n}$, and let $w = w_1w_2\dots w_n \in \mathbb{Z}_2^n$. Then $w \in C$ if and only if there exists $x = x_1x_2\dots x_m \in \mathbb{Z}_2^m$ such that $xG = w$. This is a system of $n$ equations in the variables $x_1, \dots, x_m$ and $w_1, \dots, w_n$. If we eliminate $x_1, \dots, x_m$ from this system, we obtain a homogeneous system of $n-m$ equations in variables $w_1, \dots, w_n$. This system can be written as $Hw^T = 0$, where $H$ is an $(n-m) \times n$ matrix whose $(i,j)$-entry is the coefficient of $w_j$ in the $i$-th equation. So we have $w \in C$ if and only if $Hw^T = 0$, and hence $H$ is a parity check matrix for $C$.

Conversely, if we are given a parity check matrix $H = [b_{ij}]_{(n-m)\times n}$, then for every $w \in C$ we have

$$Hw^T = 0.$$

If $w = w_1w_2\dots w_n$, then the above equality can be written as a system of $n-m$ equations in $n$ variables $w_1, \dots, w_n$. We can solve this system for $w_1, \dots, w_n$. Since the number of variables is greater than the number of equations, $n-(n-m) = m$ parameters $x_1, \dots, x_m$ will appear; in other words we obtain the solution in the form

$$w_j = a_{1j}x_1 + \ldots + a_{mj}x_m \quad (1 \le j \le n).$$

Thus, if we define $G = [a_{ij}]_{m\times n}$, we have $xG = w$, and $G$ is a generator matrix for $C$.

**Example 3.7.** Let us find a parity check matrix for the code given in Example 3.5. So we consider the system $xG = w$, where $x = x_1x_2x_3$ and $w = w_1w_2w_3w_4w_5w_6$. In expanded form this system is:

$$\begin{array}{ccccccc} x_1 & & & + & x_3 & = & w_1 \\ & & x_2 & & & = & w_2 \\ x_1 & + & x_2 & + & x_3 & = & w_3 \\ x_1 & & & & & = & w_4 \\ & & x_2 & & & = & w_5 \\ & & x_2 & + & x_3 & = & w_6. \end{array}$$

Substituting any values in for $x_1$, $x_2$ and $x_3$ would yield a codeword $W$. Instead we eliminate $x_1, x_2, x_3$ by using $x_1 = w_4$, $x_2 = w_2$, $x_3 = w_1 + w_4$ (remember, over $\mathbb{Z}_2$

we have $x_1 + x_1 = 0$). This gives us a set of 3 equations that do not involve $x_1$, $x_2$ or $x_3$:

$$\begin{array}{ccccccccccccc} w_1 & + & w_2 & + & w_3 & & & & & & & = & 0 \\ & & w_2 & & & & & + & w_5 & & & = & 0 \\ w_1 & + & w_2 & & & + & w_4 & & & + & w_6 & = & 0. \end{array}$$

Any set of 6 variables $w_1w_2w_3w_4w_5w_6$ that satisfy these three equations is a codeword. We can write these three equations as $Hw^T = 0$, where

$$H = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix},$$

and $H$ is a parity check matrix for $C$.

**Example 3.8.** Let a code $C \subseteq \mathbb{Z}_2^6$ be given by the parity check matrix

$$H = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

We find the corresponding generator matrix as follows. We consider the system $Hw^T = 0$, which is equivalent to

$$\begin{array}{ccccccccccccc} w_1 & + & w_2 & + & w_3 & & & & & & & = & 0 \\ w_1 & & & + & w_3 & + & w_4 & + & w_5 & + & w_6 & = & 0. \end{array}$$

Any choice of $w_1w_2w_3w_4w_5w_6$ which satisfies these two equations is a codeword. Let us therefore solve it for $w_1, w_2, w_3, w_4, w_5, w_6$, writing the parameters as $x_1, x_2, x_3, x_4$:

$$\begin{array}{ccccccccc} w_1 & = & x_1 & & & & & & \\ w_2 & = & & & x_2 & & & & \\ w_3 & = & x_1 & + & x_2 & & & & \\ w_4 & = & & & & & x_3 & & \\ w_5 & = & & & & & & & x_4 \\ w_6 & = & & & x_2 & + & x_3 & + & x_4. \end{array}$$

This solution can be written as $w = xG$, where

$$G = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

and $G$ is a generator matrix for $C$.

**Remark 3.9.** If we start with a generator matrix $G$, find the corresponding parity check matrix $H$ and then the corresponding generator matrix $G_1$, we need not have $G = G_1$. This is because a code may have several different generator matrices.

As one might expect, the parity check matrix also contains information about the error detecting and correcting capabilities of the code.

**Theorem 3.10.** *Let $C \subseteq \mathbb{Z}_2^n$ be a linear code with parity check matrix $H$. Then the minimum distance is equal to the size of the smallest set of linearly dependent columns of $H$. In particular, we can detect (respectively, correct) all errors of weight up to $k$ if and only if the size of the smallest set of linearly dependent columns of $H$ is $k+1$ (respectively, $2k+1$).*

**Proof.** By Theorem 3.2 since $C$ is linear the minimum distance is equal to the minimum weight of a non-zero code-word. Write $H = [c_1 \dots c_n]$, where the $c_i$ are columns of $H$. For a word $w = a_1 \dots a_n \in \mathbb{Z}_2^n$ we have $w \in C$ if and only if $Hw^T = 0$, i.e. if and only if $a_1c_1 + \dots a_nc_n = 0$. The word $w$ has weight $k$ if and only if exactly $k$ of $a_1, \dots, a_n$ are equal to 1. Therefore $C$ contains a word of weight $k$ if and only if $H$ has a set of $k$ linearly dependent columns. ■

## 4. Some more group theory: subgroups and cosets

In this section we recall some more elementary group theory that we will need in Section 5. In order to make the notation compatible with what follows, we shall use the additive notation for groups; in other words we shall denote the group operation by $+$, the identity element by 0, and the inverse of $x$ by $-x$.

A non-empty subset $H$ of a group $G$ is a *subgroup* if it is a group itself under the same operation. For example, if $G = \mathbb{Z}_6 = \{0, 1, 2, 3, 4, 5\}$, then $H = \{0, 2, 4\}$ is a subgroup, while $K = \{0, 1, 2, 3\}$ is not. Actually, it can be proved that $H$ is a subgroup of $G$ if and only if $H$ is closed under $+$ and under taking inverses.

Let $G$ be a group, let $H$ be a subgroup of $G$, and let $a \in G$. The *coset* of $H$ determined by $a$ is the set

$$a + H = \{a + h \ : \ h \in H\}.$$

The main properties of cosets are given in the following

**Theorem 4.1.** *Let $G$ be a group, and let $H$ be a subgroup of $G$. The cosets of $H$ satisfy the following properties.*

*(i)* $H = 0 + H$ *is a coset of itself.*

*(ii)* $a \in a + H$ *for every element* $a \in G$.

*(iii)* $|a + H| = |b + H|$ *for all* $a, b \in G$*; in other words, any two cosets of $H$ have the same number of elements.*

*(iv) Any two distinct cosets of $H$ are disjoint (i.e. their intersection is the empty set).*

*(v)* $G = \bigcup_{a \in G}(a + H)$*; in other words, $G$ is the union of all cosets of $H$.*

**Proof.** Exercise. ■

The above theorem can be summed up as follows: the cosets of a subgroup partition the group into blocks of equal size.

## 5. Decoding with coset leaders and syndromes

Let $C \subseteq \mathbb{Z}_2^n$ be an $m$-dimensional linear code, let $G$ be a generator matrix for $C$, and let $H$ be a parity check matrix for $C$. We have seen that $G$ yields an easy-to-compute encoding function $E : \mathbb{Z}_2^m \longrightarrow \mathbb{Z}_2^n$ given by $E(x) = xG$. In this section we discuss decoding.

First of all we introduce a restriction on the generator matrix $G$: we require that $G$ be in *standard form*, meaning that $G$ is written as

$$G = \begin{bmatrix} 1 & 0 & \dots & 0 & b_{11} & b_{12} & \dots & b_{1,n-m} \\ 0 & 1 & \dots & 0 & b_{21} & b_{22} & \dots & b_{2,n-m} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 & b_{m1} & b_{m2} & \dots & b_{m,n-m} \end{bmatrix}.$$

So $G$ consists of the identity matrix $I_m$, followed by an $m \times (n-m)$ matrix $B$; we write briefly $G = [I_m|B]$.

The reason for making this restriction is that it makes it easy to decode the code-words. Note that for every $x \in \mathbb{Z}_2^m$ we have

$$E(x) = xG = x[I_m|B] = [xI_m|xB] = [x|xB].$$

So every word $x \in \mathbb{Z}_2^m$ is encoded as a longer word beginning with $x$. Conversely, if a code-word $w = w_1w_2 \ldots w_n \in C$ is received, we ought to decode it as $w_1w_2 \ldots w_m$.

Another advantage of $G$ being in the standard form is that it is easy to find the corresponding parity check matrix.

**Theorem 5.1.** *Let $C \subseteq \mathbb{Z}_2^n$ be an $m$-dimensional linear code. Then $G = [I_m|B]$ is a generator matrix for $C$ if and only if the matrix $H = [B^T|I_{n-m}]$ is a parity check matrix for $C$.*

**Proof.** We show that if $G = [I_m|B]$ is a generator matrix for $C$, then $H = [B^T|I_{n-m}]$ is a parity check matrix for $C$, and leave the converse as an exercise.

We consider the system $xG = w$. In an expanded form this system is:

$$\begin{array}{lclclcl cl}
x_1 & & & & & & & = & w_1 \\
 & & x_2 & & & & & = & w_2 \\
 & & & & \ddots & & & & \vdots \\
 & & & & & & x_m & = & w_m \\
b_{11}x_1 & + & b_{21}x_2 & + & \ldots & + & b_{m1}x_m & = & w_{m+1} \\
b_{12}x_1 & + & b_{22}x_2 & + & \ldots & + & b_{m2}x_m & = & w_{m+2} \\
\vdots & & & & & & & & \vdots \\
b_{1(n-m)}x_1 & + & b_{2(n-m)}x_2 & + & \ldots & + & b_{m(n-m)}x_m & = & w_n
\end{array}$$

We solve this for $w$ by substituting $w_1 = x_1, \ldots, w_m = x_m$, yielding the system of $n-m$ equations:

$$\begin{array}{lclclclclllcl}
b_{11}w_1 & + & b_{21}w_2 & + & \ldots & + & b_{m1}w_m & + & w_{m+1} & & & = & 0 \\
b_{12}w_1 & + & b_{22}w_2 & + & \ldots & + & b_{m2}w_m & + & & w_{m+2} & & = & 0 \\
\vdots & & & & & & & & & & \ddots & & \vdots \\
b_{1(n-m)}w_1 & + & b_{2(n-m)}w_2 & + & \ldots & + & b_{m(n-m)}w_m & + & & & w_n & = & 0
\end{array}$$

This system of equations can be written as $Hw^T = 0$ where $H = [B^T|I_{n-m}]$. ■

The problem arises when we want to decode a word which is not a code-word. A reasonable approach to this is to find first the corresponding code-word, and then to decode this code-word, as explained above. This amounts to finding the error of transmission. However, since every word is a possible error, we cannot determine which error has occurred. Instead, we want to discover the error which is most likely to have occurred. Now remember that the probability of a single error is small, and certainly smaller than 0.5. This means that errors of small weights are likelier to occur than those of large weights. Consequently, for any received word we want to find the code-word closest to it (with respect to the Hamming distance).

Note that $C$, being a linear code, is certainly a subgroup of $\mathbb{Z}_2^n$. Thus we may talk about cosets of $C$ in $\mathbb{Z}_2^n$. If $C$ has dimension $m$, then $|C| = 2^m$, and so there are $2^n/2^m = 2^{n-m}$ cosets, say $C = C_1, C_2, \ldots, C_{2^{n-m}}$.

**Definition 5.2.** Let $C_i$ be a coset of $C$, and let $a \in C_i$. A *coset leader* of $C_i$ is any element $a \in C_i$ of minimal weight; in other words $a$ is a coset leader if for any other $b \in C_i$ we have $wt(a) \leq wt(b)$.

The following theorem shows how coset leaders give us a method for decoding with the desired properties.

**Theorem 5.3.** *Let $C \subseteq \mathbb{Z}_2^n$ be an $m$-dimensional linear code, let $C = C_1, C_2, \ldots, C_{2^{n-m}}$ be the cosets of $C$, and let $a_1, a_2, \ldots, a_{2^{n-m}}$ be respective coset leaders. If a word $w \in \mathbb{Z}_2^n$ belongs to the coset $C_i$ then $w + a_i$ is the code-word closest to $w$.*

**Proof.** From $w \in C_i = a_i + C$ it follows that $w = a_i + v$ for some $v \in C$. But then $w + a_i = w - a_i = v \in C$ is a code-word.

Let $u \in C$ be any other code-word, and let $b = w + u$. We have

$$b = w + u = a_i + v + u \in a_i + C = C_i.$$

Since $a_i$ is a coset leader for $C_i$, we have $wt(a_i) \leq wt(b)$, and so

$$d(w,u) = wt(w-u) = wt(b) \geq wt(a_i) = wt(w-v) = d(w,v),$$

as required. ■

This solution for the problem of decoding is still not entirely satisfactory: we have not avoided the need to store all the elements of $\mathbb{Z}_2^n$. We solve this final problem by introducing the following new concept.

**Definition 5.4.** Let $C \subseteq \mathbb{Z}_2^n$ be a linear code of dimension $m$, and let $H$ be its parity check matrix. For a word $w \in \mathbb{Z}_2^n$, its *syndrome* is the word $Hw^T \in \mathbb{Z}_2^{n-m}$.

The significance of syndromes for decoding is based on the following theorem.

**Theorem 5.5.** *Let $C \subseteq \mathbb{Z}_2^n$ be a linear code, let $H$ be its parity check matrix, and let $w_1, w_2 \in \mathbb{Z}_2^n$. Then $w_1$ and $w_2$ belong to the same coset of $C$ if and only if their syndromes are equal.*

**Proof.** ($\Rightarrow$) Assume that $w_1$ and $w_2$ belong to the same coset $a + C$ of $C$. This means that $w_1 = a + u$, $w_2 = a + v$ for some $u, v \in C$. Since $H$ is a parity check matrix for $C$ we have $Hu^T = Hv^T = 0$, and so

$$Hw_1^T = H(a^T + u^T) = Ha^T = H(a^T + v^T) = Hw_2^T.$$

($\Leftarrow$) Now assume that $Hw_1^T = Hw_2^T$. This implies that $H(w_1^T - w_2^T) = 0$, and hence $w_1 - w_2 \in C$. If we denote $w_1 - w_2$ by $u$, then we have $w_1 = w_2 + u \in w_2 + C$, and so $w_1$ and $w_2$ belong to the same coset of $C$. ■

So assume that we know coset leaders for $C$ (we show in Example 5.6 how to calculate them) and corresponding syndromes. Then we can find the coset leader for an arbitrary word just by computing its syndrome. Therefore, we have no need to store $2^n$ elements of $\mathbb{Z}_2^n$, but instead store $2^{n-m}$ coset leaders, and the same number of syndromes.

If we combine all the results from this section we obtain the following method for decoding.

**Decoding.** Let $C \subseteq \mathbb{Z}_2^n$ be an $m$-dimensional linear code, with the generator matrix $G$ in standard form, and the corresponding parity check matrix $H$. Prior to decoding do the following four steps:

1) calculate code-words of $C$;

2) find coset leaders for $C$;

3) for each coset leader calculate its syndrome;

4) Store the table of coset leaders and their syndromes: all the other codewords may be discarded.

To decode an arbitrary word $w \in \mathbb{Z}_2^n$ do the following four steps:

1) calculate the syndrome of $w$;

2) find the coset leader $a$ with the same syndrome;

3) let $v = a + w$;

4) decode $w$ as the first $m$ symbols of $v$.

**Example 5.6.** Let $C$ be the code with the generator matrix

$$G = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}.$$

The corresponding parity check matrix is

$$H = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

The elements of $C$ can be obtained as $xG$, $x \in \mathbb{Z}_2^3$. We obtain

$$C = \{000000, 100110, 010011, 001101, 110101, 101011, 011110, 111000\}.$$

Now we find coset leaders. We list the elements of $C$ in the first row of a table, and then we find an element of $\mathbb{Z}_2^n$ of minimal weight which is not listed. For example, we can take 100000 to be this element. We add this element to all the elements of $C$, and thus obtain the second row of our table. So after these two steps the table looks like this:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 000000 | 100110 | 010011 | 001101 | 110101 | 101011 | 011110 | 111000 |
| 100000 | 000110 | 110011 | 101101 | 010101 | 001011 | 111110 | 011000 |

Next, we again find an element of minimal weight not already listed, and add it to all the elements of $C$, obtaining the third row of the table. We keep doing this until we exhaust all the elements of $\mathbb{Z}_2^n$. We obtain the following table:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 000000 | 100110 | 010011 | 001101 | 110101 | 101011 | 011110 | 111000 |
| 100000 | 000110 | 110011 | 101101 | 010101 | 001011 | 111110 | 011000 |
| 010000 | 110110 | 000011 | 011101 | 100101 | 111011 | 001110 | 101000 |
| 001000 | 101110 | 011011 | 000101 | 111101 | 100011 | 010110 | 110000 |
| 000100 | 100010 | 010111 | 001001 | 110001 | 101111 | 011010 | 111100 |
| 000010 | 100100 | 010001 | 001111 | 110111 | 101001 | 011100 | 111010 |
| 000001 | 100111 | 010010 | 001100 | 110100 | 101010 | 011111 | 111001 |
| 100001 | 000111 | 110010 | 101100 | 010100 | 001010 | 111111 | 011001 |

It is easy to see that the rows of the table are cosets of $C$, and that the elements in the first column are coset leaders.

Next, for each leader we calculate its syndrome. For example

$$Ha_1^T = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, Ha_2^T = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix},$$

where $a_1$ and $a_2$ are the first two coset leaders. The complete table of coset leaders and syndromes is as follows:

| | |
|---|---|
| 000000 | 000 |
| 100000 | 110 |
| 010000 | 011 |
| 001000 | 101 |
| 000100 | 100 |
| 000010 | 010 |
| 000001 | 001 |
| 100001 | 111 |

To decode, say, the word $w = 110111$ we first compute its syndrome:

$$Hw^T = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

The corresponding coset leader is 000010, and so the correct code-word is $110111 + 000010 = 110101$. Consequently, $w$ should be decoded as 110.

We give a second example, and decode the word $w = 110001$. Again, we first compute its syndrome:

$$Hw^T = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

The corresponding coset leader is 000100, and so the correct code-word is $110001 + 000100 = 110101$. We decode $w$ as 110.

## 6. Perfect codes

We have seen that a code is a device for transfer of information, potentially capable of detecting and correcting random errors in transmission. However, these two functions of a code are in conflict with one another. For example, any one-element code $C = \{u\}$, $u \in \mathbb{Z}_2^n$ can correct all errors, but cannot carry any information. At the other extreme, the full code $C = \mathbb{Z}_2^n$ can carry a lot of information, but cannot detect any errors, let alone correct them.

Recall that $\binom{n}{i} = \frac{n!}{i!(n-i)!}$ is the number of ways of choosing $i$ objects from a set of $n$ objects. The following theorem gives an upper bound for the number of code-words in a code of specified error correcting capabilities.

**Theorem 6.1.** *For a code $C \subseteq \mathbb{Z}_2^n$ with minimum distance at least $2e+1$ we have*

$$|C| \leq 2^n / \sum_{i=0}^{e} \binom{n}{i}.$$

**Proof.** For each code-word $w \in C$ consider the 'ball' with centre $w$ and radius $e$:

$$B(w, e) = \{u \in \mathbb{Z}_2^n \, : \, d(u, w) \leq e\}.$$

We claim that for distinct $w_1, w_2 \in C$ we have

$$B(w_1, e) \cap B(w_2, e) = \emptyset.$$

Indeed, $u \in B(w_1, e) \cap B(w_2, e)$ would imply

$$d(w_1, w_2) \leq d(w_1, u) + d(w_2, u) \leq e + e < 2e + 1,$$

a contradiction.

Next note that

$$B(w,e) = \bigcup_{i=0}^{e} \{u \in \mathbb{Z}_2^n \ : \ d(w,u) = i\}.$$

For each $i$, the set $\{u \in \mathbb{Z}_2^n \ : \ d(w,u) = i\}$ consists of all the words which differ from $w$ in exactly $i$ positions; clearly there are exactly $\binom{n}{i}$ such words. Hence

$$|B(w,e)| = \sum_{i=0}^{e} \binom{n}{i}.$$

Since for all $w \in C$ the sets $B(w,e)$ are disjoint subsets of $\mathbb{Z}_2^n$ we conclude

$$2^n = |\mathbb{Z}_2^n| \geq \sum_{w \in C} |B(w,e)| = |C| \sum_{i=0}^{e} \binom{n}{i},$$

and the desired inequality follows. ■

**Definition 6.2.** A code $C \subseteq \mathbb{Z}_2^n$ is said to be *perfect* if it attains the equality in the previous theorem, i.e. if the minimum distance is $2e+1$ and $|C| = 2^n / \sum_{i=0}^{e} \binom{n}{i}$.

We now show that perfect codes exist.

**Example 6.3.** Let $H$ be any $d \times (2^d - 1)$ matrix, the columns of which are all non-zero vectors from $\mathbb{Z}_2^d$, and let $C \subseteq \mathbb{Z}_2^{2^d-1}$ be the code with parity check matrix $H$. It is obvious that every two columns of $H$ are linearly independent, and that one can find three linearly dependent columns. Therefore, by Theorem 3.10, the minimum distance is 3, and we can correct single errors. In the notation of Theorem 6.1 we have $n = 2^d - 1$ and $e = 1$. The generator matrix for $C$ has dimension $(n-d) \times n$, and so the number of code-words is

$$2^{n-d} = 2^n / 2^d = 2^n/(1+n) = 2^n / \sum_{i=0}^{e} \binom{n}{i}.$$

Therefore, $C$ is a perfect code; it is called the $(2^d - 1, 2^d - d - 1)$ *Hamming code.*

Let us, for example, construct the $(7,4)$ Hamming code, so that $d = 3$. We have the freedom of choice in which order to put the columns of $H$. With future use in mind, we opt for

$$H = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

By Theorem 5.1 the corresponding generator matrix is

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix},$$

and hence the code-words are

$$\begin{gathered} C = \{0000000, 1000011, 0100110, 0010101, 0001111, \\ 1100101, 1010110, 1001100, 0110011, 0101001, 0011010, \\ 1110000, 1101010, 1011001, 0111100, 1111111\}. \end{gathered}$$

The minimum weight of $C$ is 3, and hence the minimum distance is 3.

It is possible (and not too difficult) to show that Hamming codes are the only linear perfect codes with $e = 1$. However, there exist various non-linear perfect codes with $e = 1$.

For $e > 1$ the perfect codes are few and far between. For example, for $e = 2$, a perfect code $C \subseteq \mathbb{Z}_2^n$ must satisfy:

$$|C| = 2^n / \sum_{i=0}^{2} \binom{n}{i} = 2^{n+1}/(n^2 + n + 2).$$

So we must have $n^2 + n + 2 = 2^a$ for some $a$. If we multiply both sides by 4 and set $x = 2n + 1$, $y = a + 2$, we obtain the equation

$$x^2 + 7 = 2^y.$$

This equation is known as Nagell's equation, and its solutions are $(\pm 1, 3)$, $(\pm 3, 4)$, $(\pm 5, 5)$, $(\pm 11, 7)$ and $(\pm 181, 15)$ (this is non-trivial). Since $e = 2$ we must have $n \geq 2e+1 = 5$, and so $2n+1 = x \geq 11$. We see that the only possibilities are $n = 5$ and $n = 90$. For $n = 5$ we have the following

**Example 6.4.** Let $C = \{00000, 11111\} \subseteq \mathbb{Z}_2^5$. Here we have $|C| = 2$, $n = 5$, $e = 2$ and

$$|C| = 2 = 32/16 = 2^5/(1 + 5 + 10) = 2^5/(\binom{5}{0} + \binom{5}{1} + \binom{5}{2}).$$

So $C$ is a (not very exciting) perfect code; it is called the 5-*repetition code.*

On the other hand there is no perfect code $C \subseteq \mathbb{Z}_2^{90}$ with $e = 2$; we will show this in Corollary 6.2 in Chapter 5.

Consider now the case $e = 3$; note that here we must have $n \geq 7$. This time we have

$$|C| = 2^n / \sum_{i=0}^{3} \binom{n}{i} = 3 \cdot 2^{n+1}/(6 + 6n + 3n(n-1) + n(n-1)(n-2)).$$

If we put $m = n + 1$ we see that we must have

$$m(m^2 - 3m + 8) = 3 \cdot 2^a$$

for some $a$. We have the following two cases.

*Case 1: $m = 2^b$, $m^2 - 3m + 8 = 3 \cdot 2^c$ for some $b, c$.* We have $n = 2^b - 1 \geq 7$, so we must have $b \geq 3$. If $b \geq 4$ we have $m^2 - 3m + 8 \equiv 8 \pmod{16}$, so that $c = 3$ and $m^2 - 3m + 8 = 24$. But this last equation has no integer solutions. For $b = 3$ we have $m = 8$, $c = 4$, $n = 7$ and $|C| = 2$. The 7-repetition code $\{0000000, 1111111\}$ is an example of this situation.

*Case 2: $m = 3 \cdot 2^b$, $m^2 - 3m + 8 = 2^c$ for some $b, c$.* We must have $b \geq 2$ as if $b = 1$ we get $n = 5$. The case $b \geq 4$ is eliminated as in Case 1. For $b = 2$ we have $m = 12$ and $m^2 - 3m + 8 = 116$, which is not a power of 2. For $b = 3$ we have $m = 24$, $n = 23$. An example of such a code is given below.

**Example 6.5.** Consider the (7,4) Hamming code $H$, as defined in Example 6.3. Extend each code-word in $H$ by one component, which is equal to the sum of all the other components. Thus, for example, the code-word 1110000 is extended to 11100001, while the code-word 1101010 is extended to 11010100. The obtained code $\overline{H}$ has the minimum weight of a non-zero code-word equal to 4, as the minimum weight of $H$ is 3. Let $H^* \subseteq \mathbb{Z}_2^7$ be the code obtained from $H$ by reversing all the code-words, and let $\overline{H^*}$ be the code obtained from $H^*$ by adding to each code-word the extra component equal to the sum of all components. Finally, form a new code $\overline{C} \subseteq \mathbb{Z}_2^{24}$ as follows:

$$\overline{C} = \{(a+x, b+x, a+b+x) \;:\; a, b \in \overline{H},\ x \in \overline{H^*}\}.$$

If $A$ and $X$ are bases for $\overline{H}$ and $\overline{H^*}$ respectively, then one may prove that the set

$$\{(a,0,a) \;:\; a \in A\} \cup \{(0,b,b) \;:\; b \in A\} \cup \{(x,x,x) \;:\; x \in X\}$$

is a basis for $\overline{C}$. In particular, $|\overline{C}| = 2^{12}$. Also, one may prove that the minimal weight of a non-zero code-word in $\overline{C}$ is equal to 8.

Now, delete the last component of every vector in $\overline{C}$ to obtain a new code $C \subseteq \mathbb{Z}_{23}$. It still has $2^{12}$ code-words, and the minimum weight of a non-zero code-word is 7, so that $C$ corrects up to three errors ($e = 3$). Now we have

$$\begin{aligned} 2^n / \sum_{i=0}^{e} \binom{n}{i} &= 2^{23}/(\binom{23}{0} + \binom{23}{1} + \binom{23}{2} + \binom{23}{3}) \\ &= 2^{23}/(1+23+253+1771) = 2^{23}/2^{11} = 2^{12} = |C|, \end{aligned}$$

and $C$ is perfect! The code $C$ is called the (binary) *Golay code.*

It actually turns out that for $e > 1$ the Golay code and the repetition codes are the only perfect codes that exist.