## 1. Channel Encoding: basics

We consider now the basic problems of channel encoding. We assume that the messages to be sent are finite strings from a given finite alphabet $A$.

**Definition 1.** *Given two strings $x$ and $y$ of the same length $m$ over an alphabet $A$, the **distance** $d(x, y)$ is defined to be the number of coordinates where they differ. More precisely,*

$$\text{dist} : A \times A \to \mathbb{Z}, \qquad \text{dist}(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases}$$

*is extended to sequences of length $m$, $x\{= x_1 \cdots x_m\}$ and $y = \{y_1 \cdots y_m\}$ as $\text{dist}(x, y) = \sum_{i=1}^{m} d(x_i, y_i)$.*

It is clear that dist defines a distance in $A^m$ for any $m \in \mathbb{N}$. This distance, called Hamming distance, is essential for what follows.

**Exercise 2.** *Prove that Hamming distance is translation invariant.*

**Definition 3.** *A $q$-block code $C$ of **length** $n$ is a subset of $A^n$. $M = |C|$, the number of codewords, is the **size** of the code.*
*The **information rate** of $C$ is defined to be $\frac{\log_q(M)}{n}$.*
*The **distance** of the code is defined to be $d(C) = \min\{\text{dist}(c, c') : c, c' \in C; c \neq c'\}$.*

The channel encoding problem consists basically in the following: when a codeword $c \in C$ is sent, a word $x$ is received; we will use the notation $c \rightsquigarrow x$. $x$ may differ from $c$ because of interferences in the transmission. We want to construct codes that allow efficient transmission, that we identify with high information rates, and capability of error detection and/or correction:

**Definition 4.** *$C$ is $u$-error detecting if it is possible to identify the existence of, at least, $u$ errors in the transmission of any codeword.*
*$C$ is $t$-error correcting if it is possible to correct at least $t$ errors in the transmission.*
*$C$ is said to be exactly $u$-error detecting if it is $u$-error detecting but not $u+1$-error detecting. A similar definition applies to error correcting.*

Implicit in these definitions lies a concept of decoding:

**Definition 5.** ***Minimal Distance Decoding** consists in decoding each received $x$ as the codeword $c$ that minimizes $\text{dist}(x, c)$. In case there is more than one codeword that minimizes that distance, there is the option of not decoding (**incomplete decoding**) or of choosing one of the codewords (**complete decoding**).*

**Example 6** (Repetition Codes)**.** *A simple example of a code is given by encoding each symbol $a \in \{0, 1\}$ as the "constant" word $a \cdots a$ of length $2r + 1$, for some $r \in \mathbb{N}$. The information rate is clearly $\frac{1}{2r+1}$, while the distance is $2r + 1$ so this repetition code corrects $r$ errors.*
*In this case there are only two codewords and Minimal Distance Decoding consists simply in choosing the symbol that occurs in the majority of the entries (that is why the length is odd).*
*Of course, repetition codes may also be defined for larger alphabets $A$.*

**Example 7.** *A second example, that will be much explored later, is the following: we consider again the alphabet* $\{0,1\}$*, but we identify it with the finite field* $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$*; the message to be sent is decomposed in strings of length* 4 *and* $(x_1, x_2, x_3, x_4)$ *is encoded as* $(c_i)_{\in}\mathbb{F}_2^7$ *as*

$$\begin{cases} c_i = x_i \,\forall\, 1 \leq i \leq 4 \\ c_5 = c_2 + c_3 + c_4 \\ c_6 = c_1 + c_3 + c_4 \\ c_7 = c_1 + c_2 + c_4 \end{cases}$$

*In particular, this code has information rate* $\frac{4}{7}$*.*
*If a vector* $(u_i) \in \mathbb{F}_2^7$ *is received, the decoding procedure consists in computing*

$$\begin{cases} s_1 = u_5 - (u_2 + u_3 + u_4) \\ s_2 = u_6 - (u_1 + u_3 + u_4) \\ s_3 = u_7 - (u_1 + u_2 + u_4) \end{cases}$$

*It is clear that the vector* $(s_1, s_2, s_3)$ *is the zero vector if and only if* $(u_i)$ *is a codeword. Moreover, if it is not and we assume that an error occured in transmission in a single coordinate, we are able to correct it because if*

$$u_i = c_i + 1, \qquad u_j = c_j \,\forall\, j \neq i,$$

*each possible nonzero vector* $(s_1, s_2, s_3)$ *is obtained for each value of* $1 \leq i \leq 7$*.*
*In other words, for each vector* $(u_i)$ *there is a unique codeword at minimal distance from it and that distance is* 1 *unless, of course, if* $(u_i)$ *is a codeword.*

Let's consider the situation where one of these codes is used for transmission through a Binary Symmetric Channel with cross-over probability of error $p$:

If the repetition code described above is used with this channel, the decoding delivers a wrong codeword if and only if more than $r$ coordinates are changed. So the probability of decoding error is

$$P_{de} = \sum_{k=r+1}^{2r+1} \binom{2r+1}{k} p^k (1-p)^{2r+1-k} = \binom{2r+1}{r+1} p^{r+1} + \text{ terms with higher powers of } p.$$

On the other hand, if the code described in the second example is used, the decoding procedure delivers a wrong codeword if and only if more than 1 coordinate is changed, since the codeword chosen is at distance 0 or 1 from the received vector. So

$$P_{de} = \sum_{k=2}^{7} \binom{7}{k} p^k (1-p)^{7-k} = \binom{7}{2} p^2 + \text{ terms with higher powers of } p.$$

We now confirm that error detecting/correcting capability is closely related to the distance of the code:

**Proposition 8.** *If a block code $C$ with minimal distance $d$ is used and minimal distance decoding is applied, then*

a) *$C$ is $u$-error detecting if and only if $d > u$;*
b) *$C$ is $t$-error correcting if and only if $d \geq 2t + 1$.*

*Proof.* (**HW**). □

**Remark 9.** *It is important to clarify that error detection capability refers to detection of the existence of errors and not to the identification of the errors. In fact, if the codeword $c$ is sent and the received vector $x$ satisfies $\operatorname{dist}(x, c) = u < d$, then we know that $x$ is not a codeword and so detect the presence of an error; but minimal distance decoding may lead to correct $x$ to a different codeword $c'$ and so, in a certain sense, the true error was not detected. On the other hand, if $\operatorname{dist}(x, c) \geq d$, the received vector may be a codeword and in that case the existence of errors is not even detected.*

*In other words, if we use the code for error detection only (ie, we merely identify the existence of errors in transmission) then we are sure to detect all cases where less than $d$ errors occur; if we use it for error correction we are sure to correct all cases where less than $\frac{d-1}{2}$ errors occur.*
*In certain cases, we may use the code to simultaneously detect and correct errors: suppose that the minimal distance of the code is even, say $d = 2t + 2$, and that $x$ is received. If there exists a unique codeword $c$ at minimal distance from $x$ we decide to correct $x$ to $c$, otherwise we declare the existence of errors; in this way, if $t$ (or less) errors occur in the transmission, we decode correctly, while if $t+1$ errors occur we still detect correctly. Notice that (**HW**) if $d$ is odd this is no longer true.*

## 2. Linear Codes

The construction of simple and efficient error-correcting codes may be done using tools from Linear Algebra. We denote as $\mathbb{F}_q$ the field with $q$ elements. We will confirm later that such a field exists and is essentially unique if and only if $q$ is the power of a prime, and study the useful properties of these fields in detail. For the time being, we may restrict ourselves to the fields $\mathbb{F}_p$ with $p$ prime, which may be identified with the sets $\mathbb{Z}_{/p}$ of congruence classes of integers modulo $p$. Operations with elements of the field are simply the usual arithmetic operations on integers, disregarding all multiples of the modulus. As an illustration we include the sum and multiplication tables of $\mathbb{F}_5$:

| $+$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 2 | 3 | 4 |
| 1 | 1 | 2 | 3 | 4 | 0 |
| 2 | 2 | 3 | 4 | 0 | 1 |
| 3 | 3 | 4 | 0 | 1 | 2 |
| 4 | 4 | 0 | 1 | 2 | 3 |

| $\times$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 2 | 3 | 4 |
| 2 | 0 | 2 | 4 | 1 | 3 |
| 3 | 0 | 3 | 1 | 4 | 2 |
| 4 | 0 | 4 | 3 | 2 | 1 |

**Definition 10.** *A $q$-linear code $C$ of length $n$ is a vector subspace of $\mathbb{F}_q^n$.*

The size of $C$ is then $q^k$, where $k$ is the dimension of the subspace, while the information rate is $\frac{k}{n}$. $C$ is said to be a $[n,k]$-linear code (over $F_q$).

Channel Encoding of Linear Codes is as follows: Let $G$ be a $k \times n$ matrix over $\mathbb{F}_q$ whose rows constitute a basis for $C$. The messages to be sent are broken into $k$-dimensional blocks $u$ (that we take as vectors in $\mathbb{F}_q^k$), which are encoded as $uG$. $G$ is called a **generator matrix** for $C$. A different choice of basis gives rise to a different encoding with the same codewords. A particularly simple and useful choice of basis is to have $G = [I\, A]$ where $I$ denotes the $k$-dimensional identity matrix. In this case, $G$ is said to be in **standard form** and to give rise to **systematic encoding** because the first $k$ digits are always the message while the remaining $n-k$ are the redundancy digits. It can be shown that any $[n,k]$-linear code is equivalent, in a certain natural sense, to one with generator matrix in standard form. If $G$ is in standard form the first $k$ coordinates of codewords coincide with the source message, the remaining ones being the redundancy information needed to assure error correcting capability. This situation is generalized with the following definition:

**Definition 11.** *An **information set** for a $[n,k]$ code with generator matrix $G$ is a set of $k$ coordinates such that the corresponding columns of $G$ are linearly independent.*

Notice that the definition does not depend on the generator matrix, and that the knowledge of an information set allows to recover the original message from the corresponding codeword.

**Proposition 12.** *For a linear code $C$, $d(C) = \min\{w(c) : c \in C\}$ where $w : \mathbb{F}_q^n \to \mathbb{Z}$ is defined as $w(x) = \mathrm{dist}(x, 0)$, ie, the number of nonzero coordinates in $x$. $w(x)$ is called the **weight** of $x$.*

*Proof.* **HW** □

**Definition 13.** *Given a $[n,k]$-linear code $C$ with generator matrix $G$, a **parity-check matrix** for $C$ is a $(n-k) \times n$ matrix $H$ over $\mathbb{F}_q$ with linearly independent rows such that $GH^T = 0$ (where $H^T$ denotes the transpose of $H$).*

If $G = [I\, A]$ we may take $H = [-A^T\, I]$.

2.1. **Decoding for Linear Codes.** We now consider the problem of applying Minimal Distance Decoding to linear codes.

**Definition 14.** *The **syndrome** of $x \in \mathbb{F}_q^n$ is defined as* $\mathrm{syn}(x) = Hx^T$.

This definition is motivated by the following observation:

**Lemma 15.** $\mathrm{syn}(x) = 0$ *iff* $x \in C$, *and so* $\mathrm{syn}(x) = \mathrm{syn}(y)$ *iff* $x - y \in C$.

*Proof.* **HW** □

So the syndrome of a received vector $x$ reflects the occurrence of errors in the transmission: if $c \rightsquigarrow x = c + e$, $\text{syn}(x) = \text{syn}(e)$. $e$ is called the **error pattern**.
We notice also that there is a bijection between syndromes and cosets of $\mathbb{F}_q^n$ with respect to the subspace $C$; in particular, the number of syndromes is $|\mathbb{F}_q^n/C| = q^{n-k}$.

The previous definitions and observations lead to the following scheme of **syndrome decoding**: given a received $x$, we search for a possible codeword $c$ such that $x = c + e$; because we are assuming Minimal Distance Decoding, we want to find $e$ with the least possible weight.
We construct a **syndrome/coset** table, associating to each syndrome $s_i$ a representative $u_i$ with minimal weight of the corresponding coset; these representatives are called **coset leaders**. Given a received $x$ with $\text{syn}(x) = s_i$, decode $x$ as $x - u_i$, which is a codeword.
If, for some coset, there is more than one possible coset leader, we either make no choice and leave any word with the corresponding syndrome undecoded (incomplete decoding) or choose one possible coset leader (complete decoding).

As an immediate consequence of the definitions:

**Proposition 16.** *$C$ is $t$-error correcting iff all $v$ with $w(v) \leq t$ are coset leaders.*

And we have also

**Lemma 17.** *If $w(e) \leq \frac{d-1}{2}$ then $e$ is the unique coset leader of its coset.*

*Proof.* **HW** □

**Remark 18.** *This last statement shows that a practical way of determining a syndrome/coset table is to start by listing all vectors with weight less or equal than $\frac{d-1}{2}$ and compute their syndromes. If this process does not complete the table, we search for the remaining coset leaders among the available vectors with least possible weight.*

We are left with the problem of determining the distance of a linear code.

**Proposition 19.** *If $H$ is a parity-check matrix for $C$ then $H$ has $j$ linearly dependent columns iff there exists $c \in C$ with $w(c) \leq j$.*
*In consequence,*

$$d(C) = \min\{j : H \text{ has } j \text{ linearly dependent columns }\}.$$

*Proof.* **HW** □

The two examples given in the previous section are linear codes over $\mathbb{F}_2$.

**Example 20.** *The code in the second example is the Hamming $[7, 4]$ binary code. The generator matrix corresponding to the encoding described there is*

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

*while a parity-check matrix is*

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

*An input message* $(\ 1 \quad 0 \quad 1 \quad 0\ )$ *is encoded as*

$$(\ 1 \quad 0 \quad 1 \quad 0\ )G = (\ 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1\ ).$$

*Either by direct inspection or using Proposition 18, it is easy to see that this code has minimal distance* 3*. The nonzero syndromes are exactly the columns of* $H$*, which makes it obvious to identify the coset leaders of each one. The decoding procedure described before is nothing else but syndrome decoding.*

*If the vector*

$$u = (\ 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1\ )$$

*is received, the syndrome is*

$$Hu^t = (1\,0\,1)$$

*with coset leader*

$$(\ 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\ )$$

*and so* $u$ *is corrected as*

$$(\ 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1\ ).$$

**Example 21.** *The repetition code is also a linear code with length* $2r+1$*, dimension* 1 *and distance* $2r+1$*. The generator matrix is the* $1\times(2r+1)$ *matrix with all entries equal to* 1*. Syndrome decoding reproduces the simple decoding procedure described before.*

**Exercise 22** (**HW**)**.** *Write a parity-check matrix for the repetition code and confirm that each vector with weight less or equal than* $r$ *is a unique coset leader.*

**Example 23** (A non-binary code)**.** *Let* $C$ *be the linear* $[10, 5]$ *code over* $\mathbb{F}_3$ *with generator matrix*

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 2 & 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 0 & 0 & 2 & 1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 0 & 2 & 1 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 2 & 1 & 1 & 0 \end{bmatrix}.$$

**Exercise 24.**
a) *Encode the message* $(1, 2, 0, 2, 0)$*;*
b) *Compute the minimal distance of* $C$*;*
c) *Decode, by syndrome decoding, the output* $(1, 0, 2, 2, 0, 1, 2, 1, 0, 0)$ *(assume that errors are "rare" and so search for error patterns with small weight).*

## 2.2. **Dual Codes, Self-Orthogonal and Self-Dual Codes.**

**Definition 25.** *Given a code $C \subset \mathbb{F}_q^n$ its **dual code**, denoted $C^\perp$, is defined by the condition*

$$C^\perp = \{x \in \mathbb{F}_q^n : \sum_{i=1}^{n} x_i c_i = 0 \ \forall c \in C\}.$$

So, in the case of a linear code, the dual code is, as a subspace of $\mathbb{F}_q^n$, the orthogonal complement of $C$, with respect to the usual inner product (over the field $\mathbb{F}_q$)

$$< x, y >= \sum_{i=1}^{n} x_i y_i,$$

although, as we'll see below, the word complement is slightly misleading.

The generator and parity-check matrices of $C^\perp$ are easily deduced:

**Proposition 26.** *If $C$ has generator matrix $G$ and parity-check matrix $H$, then $H$ and $G$ are, respectively, a generator matrix and a parity-check matrix for $C^\perp$. In particular, if $C$ has dimension $k$, $C^\perp$ has dimension $n - k$.*

*Proof.* **HW** □

As expected,

**Proposition 27.** *For any linear code $C$, $\left(C^\perp\right)^\perp = C$.*

*Proof.* **HW** □

**Example 28.** *The dual code $C^\perp$ of the Hamming $[7, 4]$ binary code is a $[7, 3]$ code with minimal distance $4$ and is $1$-error correcting. We leave as an exercise the deduction of some properties of $C^\perp$:*

**Exercise 29 (HW).**
a) *Every nonzero codeword has weight $4$;*
b) *For every coordinate $i$ there exist exactly $4$ codewords with $c_i = 1$;*
c) *For every pair of distinct coordinates $i$ and $j$ there exist exactly $2$ codewords satisfying $c_i = c_j = 1$;*
d) *There are seven cosets with minimum weight $1$, and each one contains four vectors with weight $3$ and three vectors with weight $5$;*
e) *There are seven cosets with minimum weight $2$, and each one contains three vectors with weight $2$, four with weight $4$ and one with weight $6$;*
f) *The remaining coset has seven vectors with weight $3$ and one with weight $7$.*

**Exercise 30.** *Suppose that this code is used for transmission through a Binary Symmetric Channel. Compute the probability of decoding error, assuming complete decoding.*

Contrary to what happens in real or complex vector spaces, a nonzero vector $x \in \mathbb{F}_q^n$ may be orthogonal to itself. This possibility justifies the following

**Definition 31.** *A code $C$ is called **self-orthogonal** if $C \subset C^\perp$ and **self-dual** if $C = C^\perp$.*

**Exercise 32.** *Show that the dual code of the Hamming* $[7, 4]$ *binary code is self-orthogonal.*

Self-orthogonal and self-dual codes have interesting properties, namely in the case of binary codes. We state two of those results, whose proofs are left as exercises:

**Proposition 33.** *If* $C$ *is a binary self-orthogonal code then each codeword has even weight and* $C^{\perp}$ *contains the constant vector* $1 \cdots 1$.
*Moreover, if* $C$ *has a generator matrix each of whose rows has weight divisible by* 4*, then every codeword has weight divisible by* 4.

*Proof.* **HW**. □

**Proposition 34.** *A code* $C$ *over* $\mathbb{F}_3$ *is self-orthogonal if and only if the weight of each codeword is divisible by* 3.

*Proof.* **HW**. □

2.3. **Majority Logic Decoding.** We discuss briefly a variation of syndrome decoding. For simplicity, we'll restrict ourselves to the binary case leaving the possible generalization as an exercise.
Let $C$ be an $[n, k]$ binary code. Suppose $C^{\perp}$ contains $m$ vectors $y^1, \cdots, y^m$ satisfying

1 $y_1^i = 1$ for all $i \leq r$;
2 If $j \neq 1$ there exists *at most one* $y^i$ with $y_j^i = 1$.

Such a set is called a **orthogonal system** with respect to 1.
Assume now that $u$ is received and contains $t \leq m/2$ errors. We compute the list $< u, y^i >=< e, y^i >$, where $e$ is the error pattern. Then, if $u_1$ is correct ($e_1 = 0$), there will be, by condition 2 above, at most $t$ values of $i$ for which $< u, y^i > \neq 0$. On the other hand, if $u_1$ is incorrect ($e_1 = 1$), $< u, y^i >= 0$ happens at most for $t - 1$ values of $i$.
Noticing that $m - (t - 1) > t$ by our initial assumption, we may conclude that

**Lemma 35.** *Let* $v$ *be the number of values of* $i \leq m$ *for which* $< u, y^i >= 1$. *Assuming that the weight of the error pattern satisfies* $w(e) = t \leq m/2$,

a) *if* $v \leq t$ *then* $e_1 = 0$;
b) *if* $v > t$ *then* $e_1 = 1$.

**Exercise 36.** *Complete the proof of the lemma sketched above.*

This decoding procedure is called **Majority Logic Decoding**. Of course, a similar deduction may be made for other coordinates.
If $C^{\perp}$ contains an orthogonal systems of $m$ vectors with respect to each coordinate from an Information Set, it is then possible to decode correctly those coordinates, under the assumption that the error pattern satisfies $w(e) \leq m/2$. Sometimes this improves on the error-correcting capability guaranteed by syndrome decoding.
The following example is somewhat artificial and serves only as a simple illustration of the method (a more interesting example will be discussed later).

**Example 37.** *Let $C$ be the linear $[8,3]$ binary code with generator matrix*

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

*$C$ has minimal distance 3 and so syndrome decoding decodes correctly any single error.*

**Exercise 38.** *Find an orthogonal set of 4 vectors with respect to the first coordinate and decode, both by syndrome decoding and by Majority Logic decoding, the output $y = (1,1,1,1,0,0,0,0)$.*

## 3. Basic Constructions

There are several operations that allow to construct new codes from old ones. These may be useful to obtain codes that are more efficient or easy to deal with, but also to compare different codes and deduce properties of one from those of another.

3.1. **Equivalence of Codes.** Given a code $C$, we may contruct a new code $C'$ by applying a fixed permutation to the coordinates of all codewords. $C$ and $C'$ are **permutation equivalent**. More generally, for codes over a field $\mathbb{F}$, we may apply a permutation and multiply each coordinate by a nonzero element of $\mathbb{F}$. This equivalence relation on codes is called **monomial equivalence**. Obviously, equivalent codes have the same information rate and error detection and correction capability.
If $C$ is a linear code with generator matrix $G$ and parity-check matrix $H$, a monomial equivalent code will have generator matrix $GPD$ and parity-check matrix $HPD^{-1}$, where $P$ is a permutation matrix and $D$ a diagonal matrix with nonzero diagonal entries (**HW**). Not every linear code has a generator matrix in standard form, but every linear code is equivalent to a code with that property. These observations may be helpful, for example, in the computation of a parity-check matrix for a code.

The following operations for the construction of codes apply to general codes but will be described in detail for linear ones. In the following $C$ denotes a $[n,k,d]$ linear code over a field $\mathbb{F}$.

3.2. **Expurgation and Subcodes.** Expurgation refers simply to the operation of obtaining a new code form a given one by throwing away part of the codewords.
As an example, for any $1 \leq r < k$, we may construct from $C$ a $[n, k-r, d]$ code, choosing a vector $v \in C$ with $w(v) = d$ and a linear independent family of $k-r$ vectors containing $v$.
Another example of a subcode relies on the following definition:

**Definition 39.** *A vector $c \in F^n$ is **even-like** if $\sum_i c_i = 0$ and **odd-like** otherwise. If $C \subset \mathbb{F}^n$ is a code, $C_e$ is the subset of even-like vectors in $C$.*

**Proposition 40.** *If $C \subset \mathbb{F}^n$ is a $[n,k]$ code, $C_e$ is either a $[n,k]$ or a $[n,k-1]$ code.*

*Proof.* **HW**. □

**Remark 41.** *The relation of the minimal distance of $C$ with that of $C_e$, or with the one of the expurgated non-linear code $C \setminus C_e$ can not be determined by $d$ only.*

3.3. **Puncturing.** A $[n-1, k', d']$ code $C'$ may be obtained from $C$ by deleting a fixed coordinate in all vectors $c \in C$. We have $k' < k$ if and only if $C'$ has less codewords than $C$ if and only if there exist $c_1, c_2 \in C$ that differ exactly in the deleted coordinate, if and only if there exist a codeword with weight 1 whose nonzero coordinate is exactly the deleted coordinate. In particular $d = 1$ in this case. If $d > 1$, then $k' = k$ and $d' = d - 1$ if some minimum weight codeword is not zero at the deleted coordinate, $d' = d$ otherwise.
This shows that if all the minimum weight codewords of $C$ have a common zero coordinate $i$, puncturing $C$ at $i$ creates a code with better Information Rate and Relative distance. A similar construction (shortening) is described below.
This operation is generalized by puncturing $C$ at a subset $S$ of coordinates. The punctured code is denoted $C^{[S]}$.
The generator matrix of $C^{[S]}$ is obtained from that of $C$ by deleting the corresponding columns (and eliminating linear dependences on rows, if $k' < k$). The parity-check matrix is obtained with no extra computations from a generator of $C$ in standard form $[I|A]$ if $S$ is contained in the set of the last $n - k$ coordinates (**HW**) but, in general, it must be computed from scratch.

If $k < n$ and $H = [I_{n-k}|X]$ is the check-parity matrix in standard form of a $[n, k, d]$ code, deleting a column of $X$ produces a matrix $H_1$ that is a parity-check matrix for a $[n-1, k-1, d_1]$ code $C_1$, with $d_1 \geq d$.

**Exercise 42.** *Show that if $k < n$ and $H = [I_{n-k}|X]$ is the check-parity matrix in standard form of a $[n, k, d]$ code, deleting a column of $X$ produces a matrix $H_1$ that is a parity-check matrix for a $[n-1, k-1, d_1]$ code $C_1$, with $d_1 \geq d$.*
*Find an example where $d_1 > d$.*

3.3.1. *Shortening.* Fix a subset $S$ of coordinates and let

$$C(S) = \{c = (c_i) \in C : c_j = 0 \ \forall j \in S\}.$$

$C(S)$ is a subcode of $C$ (**HW**). The puncturing of $C(S)$ at $S$ is called the shortening of $C$ at $S$ and denoted $C_{[S]}$. Shortened codes are in fact the duals of punctured codes:

**Proposition 43.** *Given a $[n, k, d]$ code $C$ and a subset $S$ of coordinates,*

a) $\left(C^{\perp}\right)_{[S]} = \left(C^{[S]}\right)^{\perp}, \qquad \left(C^{\perp}\right)^{[S]} = \left(C_{[S]}\right)^{\perp}.$
b) *If $s = |S| < d$, then $C^{[S]}$ has dimension $k$ and $\left(C^{\perp}\right)_{[S]}$ has dimension $n - k - s$.*

*Proof.* (**HW**). □

3.4. **Direct Sum and $(u, u+v)$ constructions.** If for $i \in \{1, 2\}$ $C_i$ is a $[n_i, k_i, d_i]$ code over a field $\mathbb{F}$, their direct sum

$$C_1 \oplus C_2 = \{(c_1, c_2) : c_i \in C_i\}$$

is a $[n_1 + n_2, k_1 + k_2, d]$ code where $d = \min(d_1, d_2)$ (**HW**). Its generator and parity-check matrices are the block diagonal matrices

$$\begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix}, \qquad \begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix}$$

constructed from the generator and parity-check matrices of the two codes.

In the case $n_1 = n_2$ there is another related way of constructing a longer code, the $(u, u + v)$ construction:

**Proposition 44.** *Under the above conditions*

$$C = \{(u, u + v) : u \in C_1,\, v \in C_2\}$$

*is a* $[2n, k_1 + k_2, \min(2d_1, d_2)]$ *linear code with generator and parity-check matrices*

$$\begin{pmatrix} G_1 & G_1 \\ 0 & G_2 \end{pmatrix}, \qquad \begin{pmatrix} H_1 & 0 \\ -H_2 & H_2 \end{pmatrix}.$$

*Proof.* (**HW**). □

**Corollary 45.** *If* $C$ *is a* $[n, k, d]$ *binary code, and* $\mathbf{1}$ *denotes the vector with all entries equal to* 1*, then*

$$\{(c, c) : c \in C\} \cup \{(c, c + \mathbf{1}) : c \in C\}$$

*is a* $[2n, k + 1, \min(n, 2d)]$ *binary code.*

*Proof.* (**HW**). □

An important example of a family of codes that may be defined with this construction consists of the **Reed-Muller codes**. We describe here the original binary case:
We start by defining $R(0, m)$ to be the $[2^m, 1, 2^m]$ repetition code and $R(m, m)$ to be $\mathbb{F}_2^{2^m}$, which is the unique $[2^m, 2^m, 1]$ code. Then, for $0 < r < m$,

$$R(r, m) = \{(u, u + v) : u \in R(r, m - 1),\, v \in R(r - 1, m - 1)\}.$$

The parameters of the Reed Muller $R(r, m)$ codes may be deduced by induction:

**Proposition 46.** *For any* $0 \le r \le m$*,*

i) *The dimension of* $R(r, m)$ *is* $\sum_{j=0}^{r} \binom{m}{j}$*;*
ii) *The minimum distance of* $R(r, m)$ *is* $2^{m-r}$*.*

*Moreover, for any* $0 \le r \le s \le m$ $R(r, m) \subset R(s, m)$*.*

*Proof.* (**HW**). □

3.5. **Extension.** A code may be trivially extended by adding zero coordinates. A much more interesting operation is extending with an overall parity-check: given a $[n, k, d]$ code $C$, we define

$$\hat{C} = \{c = (c_1, \cdots, c_n, c_{n+1}) : (c_1, \cdots, c_n) \in C \wedge \sum_{i=1}^{n+1} c_i = 0\}.$$

$\hat{C}$ is a linear $[n+1, k, \hat{d}]$ code where $d \leq \hat{d} \leq d+1$, more precisely, $\hat{d} = d$ if $C$ contains a even-like (ie, with zero sum of coordinates) codeword of minimum weight, and $\hat{d} = d+1$ otherwise.
If $H$ is a parity-check matrix for $C$, the extended code $\hat{C}$ has parity-check

$$\hat{H} = \begin{bmatrix} 1 & \cdots & 1 & 1 \\ & & & 0 \\ & H & & \vdots \\ & & & 0 \end{bmatrix}$$

**Exercise 47.** *Prove the previous statements.*

3.6. **Decoding of Erasures.** As mentioned before, it may be useful to add to the output alphabet a special symbol (we will use ?) to denote an illegible symbol, which we will consider as an erasure. The decoding of erasures poses different problems because, contrary to what happens with errors (replacement of one symbol by another), the location of erasures is known.
A procedure for decoding in the presence of erasures is the following: suppose that a received message $r$ contains $l < d$ erasures and the set of erased coordinates is $L$; if we puncture $C$ at $L$ we get a $[n-l, k, d']$ code $C^{[L]}$ with $d' \geq d - l$. If $r$ contained also $v$ errors and $2v < d - l$ then we may decode the punctured message $r'$ to $c' \in C^{[L]}$ by syndrome decoding; but, in these conditions, there exists a unique $c \in C$ such that its puncturing $c^{[L]} = c'$ (**HW**). We deduced

**Proposition 48.** *Let $C$ be a $[n, k, d]$ code. Suppose $c \rightsquigarrow r$ and $r$ contains $l$ erasures, with location set $L$, and $v$ errors. Then, if $2v + l < d$, syndrome decoding in the punctured code $C^{[L]}$ allows to correct both errors and erasures.*

**Remark 49.** *Notice that the proof gives, in specific situations, a stronger result, as we may have $d' > d - l$.*

**Example 50.** *Let $C$ be the binary linear code with parity-check matrix*

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

*It has minimal distance 4. Suppose $r = (\ 1\ \ 1\ \ ?\ \ 1\ \ 0\ \ 1\ \ 0\ )$ is received. Assuming that the output contains at most one error, we may decode as follows:*

**Exercise 51.** a) *Puncture $C$ at the third coordinate and decode the corresponding vector $r' = (\ 1\ \ 1\ \ 1\ \ 0\ \ 1\ \ 0\ )$ as $c' = (\ 1\ \ 0\ \ 1\ \ 0\ \ 1\ \ 0\ )$.*

b) *Find the codeword $c \in C$ that projects to $c'$ either directly from a generator matrix or by computing the syndrome*

$$s = H(\ 1 \quad 0 \quad x \quad 1 \quad 0 \quad 1 \quad 0\ )^t$$

*and determining $x$ to satisfy $s = 0$.*

## 4. Supplementary Results and Problems

**Problem 52.** *Let $C$ the binary code of length 9 with parity check matrix*

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

a) *Determine the distance of $C$;*
b) *determine four distinct coset leaders and the respective syndromes;*
c) *Is* 000110011 *a codeword?*
d) *Decode the words* 110101101 *and* 111111111.

**Problem 53.** *Determine a parity check matrix for a binary code that has the following coset leaders:*

$$000000, 100000, 010000, 001000, 000100, 000010, 000001, 110000.$$

**Problem 54.** *Let $C$ be a binary self-dual code.*

a) *Show that if $w(x) \equiv w(y) \equiv 0 \mod 4$ then $w(x+y) \equiv 0 \mod 4$;*
b) *show that either $w(c) \equiv 0 \mod 4$ for every $c \in C$, or $w(c) \equiv 0 \mod 4$ for exactly half of the codewords;*
c) *show that the vector with every coordinate equal to 1 belongs to $C$.*

**Problem 55.** *Suppose that $u$ is a coset leader for some linear code and that $v$ has the property that $v_i \neq 0 \implies v_i = u_i$. Then $v$ is a coset leader. In particular, if there exists a coset leader with weight $w$, there exists also a coset leader with weight $w'$ for any $w' < w$.*
***Hint:*** *It is enough to suppose that $v$ and $u$ disagree in a single coordinate.*

**Problem 56.** a) *Show that, for $q = 2$ or $q = 3$, every codeword $c$ of a self-orthogonal code over $\mathbb{F}_q$ satisfies $w(c) \equiv 0 \mod q$;*
b) *Construct a self-orthogonal code over $\mathbb{F}_5$ such that at least one the codewords has weight not divisible by 5;*
c) *Show that if $x$ and $y$ are codewords in a binary self-orthogonal code with $w(x) \equiv w(y) \equiv 0 \mod 4$, then also $w(x+y) \equiv 0 \mod 4$.*

**Example 57** (Binary Golay Code)**.** *One of first important examples of error correcting codes were the* ***Golay*** *codes, of which we present an example. The binary* $[24, 12, d]$ *Golay code* $C$ *has generator matrix* $[I_{12}|A]$ *where* $I_{12}$ *denotes the* $12 \times 12$ *identity matrix and* $A$ *is an also* $12 \times 12$ *matrix defined as follows: the first row has first entry equal to* $0$ *and all the others equal to* $1$*; all other rows have first entry equal to* $1$*; the remaining entries of* $A$ *form a* $11 \times 11$ *matrix* $A'$ *with first row*

$$( \begin{array}{cccccccccccc} 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \end{array} )$$

*and whose other rows are obtained from this one by cyclically shifting to the left: the second row of* $A'$ *is*

$$( \begin{array}{cccccccccccc} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \end{array} )$$

*and so on.*

**Problem 58.** i) *Prove that* $C$ *is self-dual.*
ii) *Prove that* $d = 8$*;*

**Problem 59.** *Suppose that* $C$ *is a binary* $[31, 22, 5]$*-linear code.*

a) *Determine the number of cosets of* $C$ *and the number of coset leaders with weight* $0$*,* $1$ *and* $2$*;*
b) *Determine, for each coset, an upper bound for the number of words of weight* $3$ *contained in it;*
d) *Show that the previous computations lead to a contradiction.*

**Problem 60.** *Let* $C$ *be the code over* $\mathbb{F}_3$ *with parity-check matrix*

$$H = \begin{bmatrix} 2 & 1 & 2 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 2 & 0 & 1 & 0 & 0 \\ 2 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 2 & 1 & 2 & 0 & 0 & 0 & 1 \end{bmatrix}$$

*and let*

$$r = ( \begin{array}{cccccccccccc} 1 & 0 & 0 & ? & 0 & 0 & 2 & 0 \end{array} )$$

*be a received vector. Decode* $r$*, assuming it contains at most one error.*

## 4.1. **Hamming Codes.**

**Definition 61.** *A* ***binary Hamming code*** *is a linear code that has a parity-check matrix whose columns are all non-zero vectors in* $\mathbb{F}_2^m$ *for some* $m \in \mathbb{N}$*.*

As a direct consequence of the definition, the binary Hamming code is a $[2^m - 1, 2^m - 1 - m, 3]$-code and so a 1-error- correcting code.

**Definition 62.** *A* $t$*-error-correcting code* $C \subset \mathbb{F}_q^n$ *is* ***perfect*** *if, for every vector* $u \in \mathbb{F}_q^n$ *there exists a (necessarily unique) codeword* $c \in C$ *such that* $\operatorname{dist}(u, c) \leq t$*.*

**Lemma 63.** *Binary Hamming codes are perfect.*

*Proof.* **HW** □

**Definition 64.** *An extended binary Hamming code is obtained by extending the parity-check matrix $H$ of an Hamming code with a zero column and then with a row of all* 1*.*

The extended binary Hamming code is then a $[2^m, 2^m - 1 - m, 4]$-code (**HW**).

**Definition 65.** *A **biorthogonal** code (also called a **simplex** code) is the dual of an extended Hamming code.*

For a given $m$, the corresponding biorthogonal code is a $[2^m, m + 1, 2^{m-1}]$-code (**HW**).

$q$-ary Hamming codes (and their duals) are defined in a similar way: each non zero $v \in \mathbb{F}_q^m$ defines a 1-dimensional subspace.

**Definition 66.** *A $q$-ary Hamming code is defined by a parity-check matrix $H$ whose columns are representatives of the different* 1*-dimensional subspaces of $\mathbb{F}_q^m$.*

**Proposition 67.** *A $q$-ary Hamming code, for a given $m$, is a $[\frac{q^m-1}{q-1}, \frac{q^m-1}{q-1} - m, 3]$-code.*
*It is a perfect, exactly* 1*-error correcting code.*

*Proof.* **HW** □

**Problem 68.** *Compute the dimension and minimal distance of the the dual of a $q$-ary Hamming code.*
***Hint:*** *Use the form of the generator matrix $H$ to compute the number of zero entries in a given codeword $uH$.*

4.2. **Burst errors.** In general, we are considering the errors occuring in the transmission of a codeword to be random, as is in fact implied by our model of channel. However, it may be interesting to consider the capability of codes to detect and/or correct errors of a special kind.

**Definition 69.** *A $b$-burst is a vector with non-zero entries concentrated in $b$ consecutive positions.*

Burst error patterns are, in many cases, more likely to occur than general random errors. Although we won't try to model a channel with this property, we list a couple of related statements:

**Proposition 70.** *Let $C$ be a $[n, k]$ linear code over $\mathbb{F}_q$. If $b > n-k$ then $C$ contains a $t$-burst for some $t \leq b$.*

*Proof.* (**HW**): fix any set $S$ of consecutive $b$ coordinates and count the number of bursts with non-zero entries in $S$; show that two of them must be in the same coset. □

**Corollary 71.** *If $b > n - k$, there are $t$-bursts (with $t \leq b$) error patterns that are undetected by $C$.*

**Proposition 72.** *If $b \leq n-k$, there exists a $[n,k]$ code $C$ that detects all $t$-bursts for $t \leq b$.*

*Proof.* Let the parity-check matrix $H$ of $C$ be defined as follows: the first row is the concatenation of $\lfloor \frac{n}{b} \rfloor$ blocks of length $b$ of the form $10\cdots 0$ followed by a block of the same form to complete a vector of length $n$; the remaining $n-k-1$ rows are consecutive shifts of the first to the right.
Verify (**HW**) that any vector $v$ that is a $t$-burst with $t \leq b$ satisfies $Hv^T \neq 0$. □

**Proposition 73.** *If $C$ is a $[n,k]$ linear code that alows the correction (under minimal distance decoding) of all $t$-bursts with $t \leq b$, then $2b \leq n-k$.*

*Proof.* (**HW**). □