*Notes written by Lukasz Zaluski*

# Linear codes

**Definition 1** Let $\mathbb{F}_q$ be the finite field with $q$ elements. An $[n, k]$ *Linear Code* $\mathcal{C}$ is a subspace of dimension $k$ of the vector space $\mathbb{F}_q^n$.

Elements of the code are called *codewords*. Each message is represented as a vector of $\mathbb{F}_q^k$ and mapped to a unique codeword. The parameter $n$ is the *code length*, $k$ is the *code dimension* and the difference $n - k$ is the *redundancy* of the code. The ratio $R = k/n$ is known as *code rate* and measures the information rate, i.e. the proportion of useful (non-redundant) data transmitted in each codeword.

Codes are usually studied in the context of the Hamming metric, determined by the distance defined below.

**Definition 2** Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathcal{C}$ be two codewords. The *Hamming Distance* $\mathsf{d_H}(\boldsymbol{x}, \boldsymbol{y})$ between the codewords is the number of positions in which they differ, that is

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

It is easy to see that $\mathsf{d_H}$ is non-negative, symmetric and sub-additive, hence it is effectively a distance.

**Definition 3** Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{C}$ be a codeword. The *Hamming Weight* $\mathsf{wt_H}(\boldsymbol{x})$ of the codeword is the number of non-zero positions, that is:

$$\mathsf{wt_H}(\boldsymbol{x}) = |\{i : x_i \neq 0, 1 \leq i \leq n\}|. \tag{2}$$

Clearly, the Hamming distance and the Hamming weight define each other in the sense that $\mathsf{wt_H}(\boldsymbol{x}) = \mathsf{d_H}(\boldsymbol{x}, 0)$ and $\mathsf{d_H}(\boldsymbol{x}, \boldsymbol{y}) = \mathsf{wt_H}(\boldsymbol{x} - \boldsymbol{y})$.

For simplicity, we will denote the Hamming distance and weight by, respectively, $\mathsf{d}$ and $\mathsf{wt}$.
The following is a very important concept for linear codes.

**Definition 4** Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. The *Minimum Distance* $d$ of $\mathcal{C}$ is the minimum of the distances among all the codewords, that is

$$d = \min\{\mathsf{d}(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}, \boldsymbol{x} \neq \boldsymbol{y}\}. \tag{3}$$

**Theorem 1** *Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$ having minimum distance $d$. Then $\mathcal{C}$ is able to correct at most $w = \lfloor \frac{d-1}{2} \rfloor$ errors.*

# Bases

Linear codes can be efficiently described by matrices.

**Definition 5** Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. Let $\mathcal{B} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ be a basis for the vector subspace determined by $\mathcal{C}$. The $k \times n$ matrix $G$ having the vectors of $\mathcal{B}$ as rows is called *Generator Matrix* for $\mathcal{C}$, that is

$$G = \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_k \end{pmatrix}. \tag{4}$$

The matrix $G$ *generates* the code as a linear map: for each message $\boldsymbol{m} \in \mathbb{F}_q^k$ we obtain the corresponding codeword $\boldsymbol{m}G$. Of course, since the choice of basis is not unique, so is the choice of generator matrix. More specifically, given a generator matrix $G$, then the matrix $SG$, where $S$ is any invertible matrix, generates the same code. It is possible to choose $S$ in a particular way, so that $G = (I_k|M)$. This is called *systematic form* of the generator matrix.

Note that using a generator matrix in systematic form each message appears in the first $k$ positions of the corresponding codeword (i.e. the first $k$ positions carry the *information symbols*).

### Code equivalence

**Definition 6** Two linear codes $\mathcal{C}$ and $\mathcal{C}'$ are said to be equivalent if one is the image of the other via an isometry (i.e. a mapping that preserves distances). In particular, there are three main flavors of code equivalence:

- **Linear Isometry** an isometry $\pi : \mathbb{F}_q^n \to \mathbb{F}_q^n$ such that $\pi(u + v) = \pi(u) + \pi(v)$ and $\pi(\alpha u) = \alpha\pi(u)$
- **Semi-Linear Isometry** an isometry $\pi : \mathbb{F}_q^n \to \mathbb{F}_q^n$ such that $\pi(u + v) = \pi(u) + \pi(v)$ and $\pi(\alpha u) = \gamma(\alpha)\pi(u)$ for a field automorphism $\gamma$
- **Permutation** a permutation $\sigma \in S_n$. In this case the operation is easily representable in terms of permutation matrices. If $G$ is a generator matrix for $\mathcal{C}$ and $P$ is a permutation matrix, then a generator matrix for $\mathcal{C}'$ is $G' = GP$

**Problem 1 (Code Equivalence)** Given two generator matrices $G_1$ and $G_2$, decide whether the codes generated by them are equivalent.

This problem was introduced by Petrank and Roth, who proved that it is harder than the graph isomorphism problem but not NP-complete unless P=NP.

**Definition 7** Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. The *Dual Code* of $\mathcal{C}$ is the set $\mathcal{C}^\perp = \{\boldsymbol{x} \in \mathbb{F}_q^n : \boldsymbol{x} \cdot \boldsymbol{y} = 0 \, \forall \boldsymbol{y} \in \mathcal{C}\}$.

**Theorem 2** *Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. Then the dual code $\mathcal{C}^\perp$ is an $[n, n-k]$ linear code. Moreover, if $G = (I_k|M)$ is a generator matrix in systematic form for $\mathcal{C}$, then $H = (-M^T|I_{n-k})$ is a generator matrix for $\mathcal{C}^\perp$.*

The matrix $H$ is a very important matrix for the code $\mathcal{C}$ itself.

**Definition 8** Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$ and let $\mathcal{C}^\perp$ be its dual code. The $(n-k) \times n$ generator matrix $H$ is called *Parity-Check Matrix* for $\mathcal{C}$.

The parity-check matrix describes the code as follows:

$$\forall \boldsymbol{x} \in \mathbb{F}_q^n,\ \boldsymbol{x} \in \mathcal{C} \iff H\boldsymbol{x}^T = 0. \tag{5}$$

The name comes from the first, somewhat crude method for error detection, the *parity check*, in which a single redundancy bit is added at the end of a codeword, the bit being a 0 if the codeword has an even number of 1's and a 1 otherwise. In this way, if the received word has an odd number of 1's, it is sure that at least an error has occurred.
The vector $H\boldsymbol{x}^T$ is called *syndrome* of $\boldsymbol{x}$, and gives it name to a very efficient error-correcting method, known as *syndrome decoding*. This works by splitting the code $\mathcal{C}$ in $q^{n-k}$ cosets and then pre-computing a table containing the syndromes of all the corresponding coset leaders (that is, the minimal weight elements for each coset).

**Table 1:** Syndrome decoding

| | |
|---|---|
| Input | An $(n-k) \times n$ parity-check matrix $H$ and the received word $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{e} \in \mathbb{F}_q^n$. |
| Output | The codeword $\boldsymbol{x}$. |
| 1. | Calculate the syndrome $\boldsymbol{s} = H\boldsymbol{z}^T$. |
| 2. | Find the coset leader $\boldsymbol{\ell}$ associated to $\boldsymbol{s}$. |
| 3. | If $\boldsymbol{\ell}$ is found, return $\boldsymbol{x} = \boldsymbol{z} - \boldsymbol{\ell}$, else return $\perp$. |

This method succeeds as long as $w = \mathsf{wt}(\boldsymbol{e})$ is within the correcting radius of the code, i.e. $w \leq \lfloor \frac{d-1}{2} \rfloor$, where $d$ is the minimum distance of the code. In fact, since $\boldsymbol{x}$ is a codeword, we have $H\boldsymbol{z}^T = H\boldsymbol{x}^T + H\boldsymbol{e}^T = 0 + H\boldsymbol{e}^T = H\boldsymbol{e}^T$ and, because its weight is within the correcting radius, $\boldsymbol{e}$ is a uniquely determined coset leader. It is then easy to find the corresponding syndrome on the table.

# Special Families of Codes

A special subfamily of linear codes is that of cyclic codes.

**Definition 9** Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. We call $\mathcal{C}$ *Cyclic* if

$$\forall \boldsymbol{a} = (a_0, a_1 \ldots, a_{n-1}),\ \boldsymbol{a} \in \mathcal{C} \Longrightarrow \boldsymbol{a}' = (a_{n-1}, a_0 \ldots, a_{n-2}) \in \mathcal{C}. \tag{6}$$

Clearly, if the property holds, then all the right shifts, for any number of positions, have to belong to $\mathcal{C}$ as well.

An algebraic characterization can be given in terms of polynomial rings. In fact, it is natural to build a bijection between cyclic codes and ideals of the polynomial ring $\mathbb{F}_q[x]/(x^n - 1)$. We identify the vector $(a_0, a_1 \ldots, a_{n-1})$ with the polynomial $a_0 + a_1 x + \cdots + a_{n-1}x^{n-1}$, and then the right shift operation corresponds to the multiplication by $x$ in the ring.
Each ideal is generated by a certain polynomial $g(x)$ (for simplicity, we assume always $g$ to be monic) such that $g(x)$ divides $x^n - 1$. To each polynomial corresponds a distinct cyclic code, and we therefore call $g$ the *generator polynomial* of the code.
Like before, we can produce a generator matrix: this will have a special form.

**Definition 10** Let $\mathcal{C}$ be an $[n, k]$ cyclic code over $\mathbb{F}_q$. Then $\mathcal{B} = \{g(x), xg(x), \ldots, x^{k-1}g(x)\}$ is a basis for $\mathcal{C}$ and we obtain the generator matrix

$$G = \begin{pmatrix} g(x) \\ xg(x) \\ \vdots \\ x^{k-1}g(x) \end{pmatrix}. \tag{7}$$

Note that $G$ will be in *circulant* form, where the $i$-th row corresponds to the cyclic right shift by $i$ positions of the first row.

Generalizations include *constacyclic* codes, where in Equation (6) $\boldsymbol{a}'$ changes to $(\gamma a_{n-1}, a_0 \ldots, a_{n-2})$ for a certain constant $\gamma \in \mathbb{F}_q$, and in particular the special case of $\gamma = -1$ (*negacyclic* codes). Another generalization, the *quasi-cyclic* codes, we will see in the next lessons.

Among cyclic codes are some important families of codes such as Hamming codes, Quadratic-residue codes and especially BCH codes (Hocquenghem, Bose and Ray-Chaudhuri), which we will present briefly.

**Definition 11** A *Hamming code* is an $[n, k]$ linear code such that $k = n - r$, $n = q^r - 1/q - 1$ and parity-check matrix $H$ whose columns are 2 by 2 linearly independent.

Hamming codes attain a very important notion, that of *perfect code*.

**Definition 12 (Perfect Codes)** Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$ with minimum distance $d$. Then $\mathcal{C}$ is said to be perfect if for every possible word $\boldsymbol{y} \in \mathbb{F}_q^n$, there is a unique codeword $\boldsymbol{y} \in \mathcal{C}$ such that $\mathsf{d}(\boldsymbol{x}, \boldsymbol{y})$ is at most $e = \lfloor \frac{d-1}{2} \rfloor$.

In other words, this corresponds to saying that the $n$-dimensional balls of radius $e$ centered in the codewords of $\mathcal{C}$ "fill the space".

**Theorem 3 (Hamming Bound)** *Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. Then*

$$q^{n-k} \geq \sum_{i=0}^{e} \binom{n}{i} (q-1)^i. \tag{8}$$

**Corollary 1** *Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. Then*

$$q^{n-k} = \sum_{i=0}^{e} \binom{n}{i} (q-1)^i \iff \mathcal{C} \text{ is a perfect code.} \tag{9}$$

Hamming codes enjoy a special dedicated decoding procedure, capable of detecting 2 errors and correcting 1. This is described in the next table.

**Table 2:** Hamming Codes decoding

| | |
|---|---|
| Input | An $(n-k) \times n$ parity-check matrix $H$ and a received word $\boldsymbol{z}$. |
| Output | The codeword $\boldsymbol{x}$. |
| 1. | Calculate the syndrome $\boldsymbol{s} = H\boldsymbol{z}^T$. |
| 2. | If $\boldsymbol{s} = 0$ then return $\boldsymbol{x} = \boldsymbol{z}$. |
| 3. | Else compare $\boldsymbol{s}$ with the columns of $H$: if $\exists i : \boldsymbol{s}^T = \alpha h_i$ then set $\boldsymbol{e} = (0, \dots 0 \underbrace{\alpha}_{\text{i}} 0 \dots 0)$. Return $\boldsymbol{x} = \boldsymbol{z} - \boldsymbol{e}$. |
| 4. | Else return $\perp$. |

The next family we will look at is that of BCH codes.

**Definition 13** Let $q$ be a prime power, $b, \delta \leq n$ positive integers with $(q, n) = 1$, $m$ the multiplicative order of $q$ modulo $n$ and $\alpha$ a primitive $n$-th root of unity in $\mathbb{F}_{q^m}$. The *BCH Code* over $\mathbb{F}_q$ of length $n$ and designated distance $\delta$ is the cyclic code generated by $g(x) = \text{lcm}\{m_i(x) : b \leq i \leq b + \delta - 2\}$, where $m_i(x)$ is the minimal polynomial of $\alpha^i$ over $\mathbb{F}_q$.

If $b = 1$ then the code is said to be *narrow-sense*, and if the length is exactly $n = q^m - 1$ the code is called *primitive*.

The following is known as *BCH Bound*.

**Proposition 1** *Let $\mathcal{C}$ be a BCH code with designated distance $\delta$. Then $\mathcal{C}$ has minimum distance at least $\delta$.*

BCH codes enjoy many dedicated decoding algorithms, the most famous being probably the Berlekamp-Massey algorithm. They are appreciated for their ease of use, resulting in many applications such as satellite communications, DVD's, two-dimensional bar codes etc.
A subclass of BCH codes is of particular interest to us.

**Definition 14** Let $q$ be a prime power and $k$ a positive integer. A *Reed-Solomon (RS) Code* is a BCH code having length $n = q - 1$ and designated distance $\delta = n - k + 1$.

Again, if $b = 1$ we talk about *narrow-sense* Reed-Solomon codes.
Narrow-sense RS codes admit an alternative definition in terms of polynomial evaluation.

**Definition 15** Let $q$ be a prime power and $k$ a positive integer. Let $\mathbb{P}_k$ be the set of polynomials of degree $\leq k$ over $\mathbb{F}_q$, $\alpha$ a primitive $n$-th root of unity in $\mathbb{F}_q$ and $n = q - 1$. Then the code $\mathcal{C} = \{(f(1), f(\alpha), \dots, f(\alpha^{q-2})) : f \in \mathbb{P}_k\}$ is the narrow-sense $[n, k, n - k + 1]$ RS code over $\mathbb{F}_q$.

It is straightforward to see that the two definitions are equivalent. The above can be further generalized (Kasami, Lin and Peterson) to define an even more important family of codes.

**Definition 16** Let $q$ be a prime power and $n, k$ positive integers such that $1 \leq k \leq n \leq q$. Let $m$ be the multiplicative order of $q$ modulo $n$, $\alpha$ a primitive $n$-th root of unity in $\mathbb{F}_{q^m}$ and $\mathbb{P}_{m,k}$ be the set of polynomials of degree $\leq k$ over $\mathbb{F}_{q^m}$. Fix distinct $\boldsymbol{x} = (x_1, \dots, x_n)$ and non-zero $\boldsymbol{y} = (y_1, \dots, y_n)$ in $\mathbb{F}_{q^m}^n$. Then the *Generalized Reed-Solomon (GRS) Code* of order $r = n - k$ is the code $\mathsf{GRS}_r(\boldsymbol{x}, \boldsymbol{y}) = \{(y_1 f(x_1), y_2 f(x_2), \dots, y_n f(x_n)) : f \in \mathbb{P}_{m,k}\}$.

Clearly, the narrow-sense RS code $\mathcal{C}$ defined above is the GRS code $GRS_r(\boldsymbol{x}, \boldsymbol{y})$ having $m = 1$, $n = q - 1$, $x_i = \alpha^{i-1}$ and $y_i = 1$ for all $i = 1, \ldots, n$.

**Theorem 4 (Singleton Bound)** *Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$ with minimum distance $d$. Then*

$$k \leq n - d + 1. \tag{10}$$

**Definition 17** Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. Then $\mathcal{C}$ is said to be *maximum distance separable (MDS)* if it satisfies the Singleton bound, i.e $k \leq n - d + 1$.

GRS codes have the important property of being MDS, since their minimum distance is exactly $n - k + 1$. Moreover, it is possible to prove (for example, see book by MacWilliams and Sloane) that $\mathsf{GRS}_r(\boldsymbol{x}, \boldsymbol{y})^{\perp} = \mathsf{GRS}_{n-r}(\boldsymbol{x}, \boldsymbol{y}')$ for a certain sequence $\boldsymbol{y}' \in \mathbb{F}_{q^m}$. With the canonical choice of basis $(1, x, \ldots, x^{k-1})$ we can describe the generator matrix of the dual, that, as we know, is a parity-check matrix for $\mathsf{GRS}_r(\boldsymbol{x}, \boldsymbol{y})$, in the following form:

$$H(\boldsymbol{x}, \boldsymbol{y}') = \begin{pmatrix} y'_1 & \cdots & y'_n \\ y'_1 x_{frm[o]--} & \cdots & y'_n x_n \\ \vdots & \vdots & \vdots \\ y'_1 x_{frm[o]--}^{r-1} & \cdots & y'_n x_n^{r-1} \end{pmatrix}. \tag{11}$$

It is then possible to describe GRS codes through the above parity-check matrix (for ease of notation, we swap the roles of $\boldsymbol{y}$ and $\boldsymbol{y}'$).

**Definition 18** Let the integers $q, m, n, k$, the field element $\alpha$ and the sequences $\boldsymbol{x}, \boldsymbol{y}$ be defined as above. Then $\mathsf{GRS}_r(\boldsymbol{x}, \boldsymbol{y})$ is the code with parity-check matrix $H(\boldsymbol{x}, \boldsymbol{y})$.