CODING THEORY: LECTURES 1 AND 2

ELISA LORENZO GARCÍA

Contents

1. Error Correcting Codes

I don't know how much you know about error correcting codes, but the idea is that we send extra information through a channel in such a way that even if some information is lost or changed we can still recover the message. This is what we call redundancy.

Recovering information from a partial message is for example what we do when we notice and correct misspelling. Or what people do reading Hebrew: only the consonants are written and the vowels are left out.

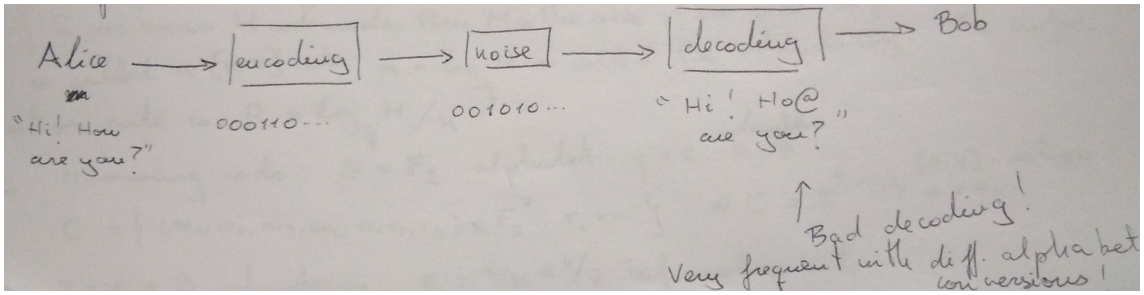
The science of deleting redundant information to store it in less memory is called data compression. The idea in error-correcting codes is the converse. One adds redundant information in such a way that it is possible to detect or even correct errors after transmission.

Example 1.1. radio contacts between pilots: A of Alpha, B of Bravo, C of Charlie, D of ...

Example 1.2. Parity check symbol. Spanish DNI: 05306738V (remainder by 23 and read a table).

Legend says that Hamming was so frustrated the computer halted every time it detected a error after he handed in a stack of punch cards, he thought about a way the computer would be able not only to detect the error but also to correct it automatically. Hamming theory is about the actual construction, the encoding and decoding of codes and uses tools from combinatorics, algebra and geometry. Shannon leads to information theory: probability sense.

Let's formalize a little bit more those ideas:



Example 1.3. Repetition Code. $Hi! \to HHii!!$ if $HHii!?$ at least one error, but you cannot correct.

$Hi! \to HHHiii!!!$ if $HHHiii!!?$, one error and I can correct it. But not two.

The information rate is $1/2$ and $1/3$ respectively.

Example 1.4. Parity check. The message $(m_1, m_2, m_3, m_4) \in \mathbb{F}_2^4$. Redundancy: $r_1 = m_1 + m_2$, $r_2 = m_3 + m_4$, $r_3 = m_1 + m_3$, $r_4 = m_2 + m_4$, $r_5 = m_1 + m_2 + m_3 + m_4$.

$$\begin{matrix} m_1 & m_2 & r_1 \\ m_3 & m_4 & r_2 \\ r_3 & r_4 & r_5 \end{matrix}$$

The information rate is $4/9 > 1/3$, also corrects one error and up to 3 erased bits.

Example 1.5. Hamming: message $(m_1, m_2, m_3, m_4) \in \mathbb{F}_2^4$. $r_1 = m_2 + m_3 + m_4$, $r_2 = m_1 + m_3 + m_4$ and $r_3 = m_1 + m_2 + m_4$. The information rate is $4/7$ and one error may be corrected. Decoding: If the 3 equalities are true no error, if only false the first one, then r_1 is wrong; if the first and the second, then m_3; etc.

THE MAGIC TRICK

Definition 1.6. Let Q be a set of q symbols called the alphabet. Let Q^n be teh set of all n-tuples $x = (x_1, ..., x_n)$ with entries $x_i \in Q$. A block code C of length n over Q is a nonempty subset of Q^n. The elements of C are called codewords. If C contains M codewords, then M is the size of the code. If $M = q^k$, then C is called an $[n, k]$ code. $n - log_q M$ is called the redundancy. The information rate is $R = log_q M/n$.

Example 1.7. For the Hamming code, see example 1.5: $Q = \mathbb{F}_2$, $q = 2$, $n = 7$. $\#C = 2^4$ so $k = 4$ and we have a $[7, 4]$-code. The redundancy is $n - k = 3$ and the information rate is $R = k/n = 4/7$.

Definition 1.8. For $x = (x_1, ..., x_n)$, $y = (y_1, ..., y_n) \in Q^n$, the Hamming distance $d(x, y)$ is defined as $|\{i | x_i \neq y_i\}|$.

Proposition 1.9. *(1) $d(x, x) \geq 0$ and equality holds iff $x = y$.*
(2) $d(x, y) = d(y, x)$ symmetry
(3) $d(x, z) \leq d(x, y) + d(y, z)$ triangle inequality

Proof. for iii) if $x_i \neq z_i$ then $x_i \neq y_i$ or $y_i \neq z_i$. □

Definition 1.10. The minimum distance of a code $\{0\} \neq C \subseteq Q^n$ is defined as $d = d(C) = min\{d(x, y) | x, y \in C, x \neq y\}$.

Definition 1.11. A linear code C is a linear subspace of $\mathbb{F}_q^n$. We denote it by $[n, k]_q$ or $[n, k, d]_q$ where n, k, d are its parameters.

Definition 1.12. For a word $x \in \mathbb{F}_q^n$: $supp(x) = \{i | x_1 \neq 0\}$, $w(x) = \#supp(x)$ and $W(C) = min\{w(c) | c \in C, c \neq 0\}$ for $C \neq \{0\}$.

Proposition 1.13. *For a linear code $d = w$.*

Proof. $d(x, y) = d(x - x, y - x) = d(0, y - x) = w(y - x)$ □

Definition 1.14. Ball of radius r around x: $B_r(x) = \{y \in Q^n | d(x, y) \leq r\}$. The sphere: $S_r(x) = \{y \in Q^n | d(x, y) = r\}$.

Example 1.15. (Hamming, see 1.5)

m_1	m_2	m_3	m_4	r_1	r_2	r_3
0	0	0	0	0	0	0
0	0	0	1	1	1	1
0	0	1	0	1	1	0
0	0	1	1	0	0	1
0	1	0	0	1	0	1
0	1	0	1	0	1	0
0	1	1	0	0	1	1
0	1	1	1	1	0	0
1	0	0	0	0	1	1
1	0	0	1	1	0	0
1	0	1	0	1	0	1
1	0	1	1	0	1	0
1	1	0	0	1	1	0
1	1	0	1	0	0	1
1	1	1	0	0	0	0
1	1	1	1	1	1	1

$W(C) = 3$ then $d = 3$ then $[7, 4, 3]_2$-linear code. $C \cap B_2(x) = \{x\}$. There are 1 word of weight 0, 7 of weight 3, 7 of weight 4 and 1 of weight 8.

The weight enumerator polynomial

$$W_C(x, y) = \sum_{w=0}^{n} A_w x^{n-w} y^w = x^7 + 7x^4y^3 + 7x^3y^4 + y^7$$

(A_w is equal to the number of words of weight w).

Mac Willians identity $W_{C^\perp}(x, y) = q^{-k} W_C(x + (q - 1)y, x - y)$, see [, Section 3.1.3].

A code with minimal distance d can detect and correct up to $\lfloor \frac{d-1}{2} \rfloor$.

The main problem of error correcting codes from Hamming's point of view is to construct for given length and k a code with the largest possible minimal distance (+ efficient encoding and decoding).

POINTS DRAWING

Theorem 1.16. *(Berlekamp, McEliece, Van Tilborg, 1978) The following problem is NP-complete: let $C \subseteq \mathbb{F}_q^n$, $y \in \mathbb{F}_q^n$ and $t \in \{0, 1, .., n\}$. Decide if there exists a $c \in C$ such that $d(y, c) \leq t$.*

Difficult problems are good for crypto!!

2. More on linear codes

Let C be a $[n, k]$ linear code over $\mathbb{F}_q$. Let $\{g_1, ..., g_k\}$ be a basis of C. $g_i = (g_{i1}, ..., g_{in})$. Denote

$$G = \begin{pmatrix} g_1 \\ ... \\ g_n \end{pmatrix} = \begin{pmatrix} g_{11} & ... & g_{1n} \\ ... & ... & ... \\ g_{k1} & ... & g_{kn} \end{pmatrix}$$

We call it a (it is not unique) generator matrix. Every $c \in C$ can be uniquely written as $m_1 g_1 + ... + m_k g_k$ where $m_i \in \mathbb{F}_q$. Then $c = mG$ with $m = (m_1, ..., m_k)$. We have a encoding: $\mathcal{E} : C \simeq \mathbb{F}_q^k \to \mathbb{F}_q^n$.

Example 2.1. The linear code with parameters $[n, 0]$ and $[n, n]$ are the trivial codes $\{0\}$ and $\mathbb{F}_q^n$, they have the empty matrix and the $n \times n$ identity matrix as generator matrices.

Example 2.2. For the Hamming code in example 1.5, we need 4 linearly independent vectors:

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Remark 2.3. The generator matrix is not unique, but the reduced row echelon form it is:

(1) all rows with only 0's are at the bottom
(2) in every row, the first element not equal to 0 is a 1, the pivot.
(3) below and above a pivot there are only 0's
(4) the pivots form an staircase

We use Gauss elimination to compute it!

3. Parity check and dual code

There are two standard ways to describe a subspace of a linear space: explicitly (basis) or implicitly (equations).

Let C be a $\mathbb{F}_q$-linear $[n, k]$ code. Let H be an $m \times n$ matrix with entries in $\mathbb{F}_q$ and such that C is the null space of H. So C is the set of all $c \in \mathbb{F}_q$ such that $Hc^t = 0$. The m equations of H are called parity check equations. We have $k \geq n - m$. If we have $k = n - m$, H has rank $n - k$ and it's called a parity check matrix.

Remark 3.1. G is made up of a basis of the kernel of H. So $HG^t = 0 = GH^t$. From th reduced form of G is easy to compute H.

Example 3.2. Over $\mathbb{F}_3$, if $G = \begin{pmatrix} 1 & 0 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{pmatrix}$ then $H = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{pmatrix}$ (on the left we write a $n - k \times n - k$ invertible matrix and on the right a $n - k \times k$ matrix that makes H satisfies $GH^t = 0$).

Theorem 3.3. *d is the smallest integer such that d columns of H are l.d. (or equivalently, the greatest d such that all $d - 1$ columns are l.i.).*

Proof. $H = (h_1, .., h_k)$, $c \in C$, $w = w(c)$. $supp(c) = \{j_1, .., j_w\}$ with $1 \leq j_1 \leq ... \leq j_w \leq n$. $Hc^t = 0$, then $c_{j_1} h_1 + ... + c_{j_w} h_{j_w} = 0$ with $c_{j_i} \neq 0$, then h_{j_i} are l.d.

Conversely, if $h_{j_1}, ..., h_{j_w}$ are dependant there exist a_i such that $\sum a_i h_{j_i} = 0$. We take $c = (c_1, ..., c_n)$ with $c_j = 0$ if $j \neq j_i$ and $c_j = a_i$ if $j = j_i$, then $Hc^t = 0$, $c \in C$ and $w(c) = w$. □

4. The dual code

Definition 4.1. Hamming code general case: $n = \frac{q^r - 1}{q - 1}$ and $H_r(q)$ be a $r \times n$ matrix over $\mathbb{F}_q$ with non-zero columns and no two l.d. $\mathcal{H}_r(q)$ q-ary Hamming code: has $H_r(q)$ as parity check matrix. $\mathcal{S}_r(q)$ q-ary simplex code: has $H_r(q)$ as generator matrix.

Definition 4.2. C is an $[n,k]$-code, the dual code is defined by $C^\perp = \{x \in \mathbb{F}_q^n | c \cdot x = 0 \, \forall c \in C\}$. A generator matrix of $C^\perp$ is a parity check matrix of C.

Proposition 4.3. *$C^\perp$ is a $[n, n-k]$-code, and $(C^\perp)^\perp = C$.*

Proof. $HG^t = GH^t = 0$. □

For the distance there is not an easy relation.

Example 4.4. $\{0\}$ and $\mathbb{F}_q^n$ are dual codes.

Example 4.5. $\mathcal{H}_r(q)$ and $\mathcal{S}_q(r)$ are dual codes by definition.

Definition 4.6. C_1 and C_2 are orthogonal if $C_1 \subseteq C_2^\perp$ and $C_2 \subseteq C_1^\perp$.

5. Golay codes

The extended binary Golay code G_{24} is a $[24, 12, 8]$-linear code. The perfect binary Golay code G_{23} is a $[23, 12, 7]$-linear code. They differ by a parity bit. They have automorphism group the (huge) Matthieu groups M_{24} and M_{23}.

As a cyclic code G_{23} is generated by $x^{11} + x^{10} + x^6 + x^5 + x^4 + x^2 + 1$.

They were used for data transmission in the Voyager 1 and 2 by the NASA to get color pictures of Jupiter and Saturn (previous Hadamard codes were not enough, only black and white ones).

There are also a ternary Golay code with parameters $[11, 6, 5]_3$. We also have the extended one with $[12, 6, 6]_3$. By Golay in 1949 and by football pool enthusiast in 1947, both independently: 11 games, 729 bets, then one with at most 2 errors.

LA QUINIELA

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

6. Decoding and the error probability

Let C be a linear code in $\mathbb{F}_q^n$ of minimum distance d. If $c \in C$ is a transmitted codeword and r is the received word, then $e = r - c$ is error vector, $supp(e) = \{i | r_i \neq c_i\}$ is the error positions and its cardinality is $w(e)$. If $w(e) < d/2$ then the nearest codeword to r is unique.

$\mathcal{E} : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is an encoder of C and $\mathcal{D} : \mathbb{F}_q^n \to \mathbb{F}_q^k \cup \{?\}$ (? is a failure) is a decoder if $\mathcal{D}\mathcal{E} = Id$.

Brute force for decoding has a complexity of nq^k.

Definition 6.1. $s = rH^t = eH^t$ is called the syndrome of r with respect to H.

Preprocess a look-up table of pairs (s, e) gives a decoder.

6.1. **The Symmetric channel.** The q-ary symmetric channel: q-ary words are sent with independent errors and the $q-1$ wrong symbols appear with probability $\frac{p_0}{q-1}$. So p_0 is the probability of error. Moreover, we ask $P(c) \equiv \frac{1}{|C|}$ for all $c \in C$ and $P(r|c)$ only depending on $d(r, c)$.

The main problem of error-correcting codes from "Shannon's point view" is to construct efficient encoding and decoding algorithms of codes with the smallest error probability

(i.e. $1-\sum_{c\in C} P(c)\sum_{\mathcal{D}(r)=c} P(r|c)$) for a given information rate and cross-over probability p_0.

Theorem 6.2. *(Shannon's random coding theorem for a q-ary symmetric channel, 1948) The error probability vanishes for $n \to \infty$ for a fixed code rate $R < 1$.*

7. Equivalent codes

Let $M \in \mathrm{GL}(n,q)$. It defines a bijection of $\mathbb{F}_q^n$. $\#\,\mathrm{GL}(n,q) = (q^n-1)(q^n-q)...(q^n-q^{n-1})$. $\phi : \mathbb{F}_q^n \to \mathbb{F}_q^n$ is called an isometry if $d(\phi(x),\phi(y)) = d(x,y)$. The set of isometries $\mathrm{Isom}(n,q)$ is a group (isometries are not necessarily linear maps).

A monomial matrix of size n is an $n \times n$ matrix with entries in $\mathbb{F}_q$ such that every row and every column has exactly one non-zero entry.

Proposition 7.1. *Let $M \in \mathrm{GL}(n,q)$. The following are equivalent:*

(1) M is an isometry
(2) $w(M(x)) = w(x)$ for all $x \in \mathbb{F}_q^n$
(3) M is a monomial matrix

Proof. i) $\Longrightarrow$ ii) and iii) $\Longrightarrow$ i) are clear. For ii) $\Longrightarrow$ iii) take the canonical basis and its image. □

Corollary 7.2. $\mathrm{GL}(n,q) \cap \mathrm{Isom}(n,q) = \mathrm{Mono}(n,q)$

Definition 7.3. $C \equiv D$ are equivalent if there exists $\phi \in \mathrm{Isom}(n,q)$ such that $D = \phi(C)$.
$C \simeq D$ are linearly equivalent if there exists $M \in \mathrm{Mono}(n,q)$ such that $M(C) = D$.

Definition 7.4. We talk about the automorphism group and the monomial automorphism group.

Proposition 7.5. *(1) $C \equiv D$ then $C^\perp \equiv D^\perp$*
(2) $C \simeq D$ then $C^\perp \simeq D^\perp$
(3) $C \equiv D$ then $C \simeq D$
(4) $C \simeq D$ then same parameters

Example 7.6. Up to linear equivalence there is only one $[7,4,3]_2$-code: Let H a parity check matrix. Then it is a 3×7 matrix. Since the distance is 3, no column is equal to zero, no two columns are l.d., but there are 3 that they are. Then up to column permutation we have:

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix},$$

and we have the Hamming code.

8. Exercises

Exercise 8.1. Compute the parameters of the n-repetition code $C \subseteq \mathbb{F}_q^n$. How many errors can be corrected? Give the weight enumerator polynomial and the generator matrix.

Exercise 8.2. Idem for the parity check in example 1.4.

Exercise 8.3. Check that the parity check code in Example 1.4 can correct until 3 erased bit.

Exercise 8.4. Let $\#Q = q$ and $x \in Q$. Prove that $|S_i(x)| = \binom{n}{i}(q-1)^i$ and that $|B_r(x)| = \sum_{i=0}^{r} \binom{n}{i}(q-1)^i$.

Exercise 8.5. Explain the Magic Trick.

Exercise 8.6. Take the even weight code $C \subseteq \mathbb{F}_q^n$ with $n \geq 2$: all words with even weight. Compute the parameters $n, q, M, k, w, d, R = k/n$. Check that it is only a linear code for $q = 2$ and give a generator matrix.

Exercise 8.7. Check the Mac Willians identity for C the trivial codes $\{0\}$ and $\mathbb{F}_q^n$. Also for C the code in exercise 8.6 with $q = 2$.

Exercise 8.8. Take the $\mathbb{F}_5$ linear code with generator matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & 3 & 4 & 0 \\ 0 & 1 & 4 & 4 & 1 & 1 \end{pmatrix}$$

Compute the reduced row echelon form and its parameters.

Exercise 8.9. Prove that the q-ary Hamming code $\mathcal{H}_r(q)$ has parameters $[n, n-r, 3]$ for $r \geq 2$. And the q-ary simplex code $\mathcal{S}_r(q)$ is a constant weight code with parameters $[\frac{q^r-1}{q-1}, r, q^{r-1}]$.

Exercise 8.10. Prove that the binary even weight code and the repetition code are dual.

Exercise 8.11. Prove that the code over $\mathbb{F}_5$ given by the generator matrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 & 3 & 3 \\ 0 & 1 & 0 & 2 & 2 & 4 \\ 0 & 0 & 1 & 3 & 1 & 3 \end{pmatrix}$$

is self-dual.

Exercise 8.12. Give an example of a ternary $[4, 2]$ self-dual code and show that there is no ternary self-dual code of length 6.

Elisa Lorenzo García, Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France.

Email address: elisa.lorenzogarcia@univ-rennes1.fr