LECTURE 7

INTRODUCTION TO LINEAR CODES

Information Theory

Thomas Debris-Alazard

Inria, École Polytechnique

Lecture 6:

It is possible to communicate Rn bits by sending n bits through a noisy channel Q if and only if

$$R \leq C(Q) \ \left(\textit{capacity}\right)$$

$\longrightarrow$ The proof relies on the use of **block-codes**:

we encode a symbol into a block-code which adds **redundancy**

An example: spell your name over the phone, send first names!

M like Mike, **O** like Oscar, **R** like Romeo, **A** like Alpha, **I** like India and **N** like November

Block-codes reach the capacity of discrete memoryless channels, but. . .

- To encode messages to send we need to store a table of exponential size. . .
- Encoding is an issue but also decoding, *i.e.*, recovering the sent message from a noisy version $\Big($in Shannon's proof we need to compute an exponential number of probabilities$\Big)$

Our wish list: defining a sub-class of codes verifying

1. Admitting an efficient encoding algorithm
2. Admitting an efficient decoding algorithm
3. Reaching the capacity

$\longrightarrow$ Linear codes! At least they verify 1 . . .

1. Basics on Linear Codes
2. Dual Representation of Linear Codes
3. Hamming Distance/Weight
4. Bounds on Minimum Distance
5. Reed-Solomon Codes and their Decoding Algorithm

BASICS ON LINEAR CODES

A finite field $\mathbb{F}_q$ is a finite set with size q admitting operations $\left(+, -, \times, /\right)$

- We necessarily have $q = p^m$ for some prime number p and m integer > 0
- Algebraic structure: $\mathbb{F}_{p^m} = \mathbb{F}_p[X]/(P(X))$ where $P \in \mathbb{F}_p[X]$ is a polynomial of degree m and irreducible $\left(P = QR\text{, implies that } P \text{ or } Q \in \mathbb{F}_q\right)$

$$\mathbb{F}_4 = \mathbb{F}_2[X]/(1 + X + X^2)$$
$$X(1 + X) = X + X^2 = -1 = 1$$

Be careful:

$$\mathbb{F}_q = \mathbb{Z}/q\mathbb{Z} \iff q \text{ is prime}$$

An important example: the binary field $\mathbb{F}_2$

$$\mathbb{F}_2 = \{0, 1\} \text{ where}$$
$$0 + 1 = 1 + 0 = 1, 0 + 0 = 1 + 1 = 0, 1 \times 0 = 0 \times 1 = 0 \times 0 = 0 \text{ and } 1 \times 1 = 1$$

$$\mathbb{F}_q^n = \underbrace{\mathbb{F}_q \times \cdots \times \mathbb{F}_q}_{n \text{ times}} \text{ is a } \mathbb{F}_q\text{-vector space}$$

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n)$$

$$\forall \lambda \in \mathbb{F}_q, \quad \lambda \cdot (x_1, \ldots, x_n) = (\lambda x_1, \ldots, \lambda x_n)$$

Linear codes:

A linear code $\mathcal{C}$ is a subspace of $\mathbb{F}_q^n$

When $\mathcal{C}$ has dimension k, we say that it is an $[n, k]_q$-code: n **length**, k **dimension**

Linear codes are block-codes when the alphabet is a finite field + a linear structure (subspace)

First example: repetition code of length 3

$$\{(0, 0, 0), (1, 1, 1)\} \text{ is a } [3, 1]_2\text{-code}$$

Rate of linear codes:

An $[n, k]_q$-code has cardinal q^k (why?) and its rate R is equal to

$$R = \frac{\log_q q^k}{n} = \frac{k}{n}$$

Non trivial linear codes:

1. $\left\{(f(x_1), \ldots, f(x_n)) : f \in \mathbb{F}_q[X],\ \deg(f) < k\right\} \subseteq \mathbb{F}_q^n$
2. Given two linear codes $U, V \subseteq \mathbb{F}_q^{n/2}$, $\left\{(\mathbf{u}, \mathbf{u}+\mathbf{v}) : \mathbf{u} \in U \text{ and } \mathbf{v} \in V\right\} \subseteq \mathbb{F}_q^n$

Exercise Session:

What are the dimensions of the above linear codes?

How to represent an $[n, k]_q$*-code? It has size* q^k*, is a table of this size necessary?*

How to represent an $[n,k]_q$-code? It has size q^k, is a table of this size necessary?

No!

Basis/Primal representation:

An $[n,k]_q$-code $\mathcal{C}$ admits a basis $\mathbf{b}_1,\dots,\mathbf{b}_k \in \mathbb{F}_q^n$

$$\mathcal{C} = \left\{ \mathbf{mG} : \mathbf{m} \in \mathbb{F}_q^k \right\} \text{ where the rows of } \mathbf{G} \in \mathbb{F}_q^{k\times n} \text{ are the } \mathbf{b}_i\text{'s}$$

The matrix $\mathbf{G}$ is called a generator matrix of $\mathcal{C}$

Redundancy versus rate:

Given a binary code $\mathcal{C}$ of dimension k, we can easily encode k bits $(m_1,\dots,m_k)$ as $\mathbf{mG} \in \mathcal{C}$ where

$$\mathbf{G} \text{ generator matrix } \left(\text{to encode does not necessitate to store an exponential size table} \right)$$

$$\text{We have mapped } k \text{ bits to } n \text{ bits! The } \left(\text{normalized}\right) \text{ redundancy } (n-k)/n = 1 - k/n = 1 - R$$

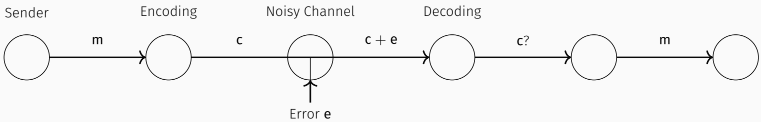
$$R \approx 0\text{: a lot of redundancy} \quad ; \quad R \approx 1\text{: few redundancy}$$

Particular case of $\mathbb{F}_2$, but can be generalized to $\mathbb{F}_q$ by encoding elements with $\log_2(q)$ bits

How to transmit k bits over a **noisy channel**?

1. **Linear code:** fix $\mathcal{C}$ subspace $\subseteq \mathbb{F}_2^n$ of dimension $k < n$
2. **Encoding:** map $(m_1, \ldots, m_k) \longrightarrow \mathbf{c} = (c_1, \ldots, c_n) \in \mathcal{C}$ task adding $n - k$ bits redundancy

   $\longrightarrow$ as $\mathcal{C}$ is linear the encoding is easy $\big($only linear algebra$\big)$, *i.e.*, $\mathbf{c} = \mathbf{mG}$
3. Send $\mathbf{c}$ across the noisy channel, errors happen and some bits of $\mathbf{c}$ are modified



Decoding:

$\longrightarrow$ from $\mathbf{c} + \mathbf{e}$: recover $\mathbf{e}$ and then $\mathbf{c}$. Now as $\mathbf{G}$ has rank k, we easily recover $\mathbf{m}$ by Gaussian elimination $\big($we use the linearity$\big)$

DUAL REPRESENTATION OF CODES

Linear codes as subspaces can also be written as the kernel of a matrix

Dual code:

Given an $[n,k]_q$-code $\mathcal{C}$, its dual $\mathcal{C}^\perp$ is an $[n,n-k]_q$-code defined as

$$\mathcal{C}^\perp = \left\{ \mathbf{c}^\perp \in \mathbb{F}_q^n : \ \forall \mathbf{c} \in \mathcal{C}, \ \langle \mathbf{c}^\perp, \mathbf{c} \rangle \overset{\text{def}}{=} \sum_{i=1}^{n} \underbrace{c_i^\perp c_i}_{\in \mathbb{F}_q} = 0 \right\}$$

Parity-check/Dual representation:

$\mathcal{C}^\perp$ is an $[n,n-k]_q$-code. Furthermore, for any generator matrix $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$ $\Big($rows of $\mathbf{H}$ form a basis of $\mathcal{C}^\perp\Big)$ we have,

$$\mathcal{C} = \left\{ \mathbf{c} \in \mathbb{F}_q^n : \ \mathbf{H}\mathbf{c}^\top = \mathbf{0} \right\}$$

Furthermore, any matrix $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$ with rank $n-k$, such that $\mathcal{C}$ is its right kernel, forms $\Big($considering its rows$\Big)$ a basis of $\mathcal{C}^\perp$.

Such matrix $\mathbf{H}$ is called a parity-check matrix of $\mathcal{C}$

Proof:

It is clear that $\mathcal{C}^\perp$ is a subspace of $\mathbb{F}_q^n$. Let us show that $\mathcal{C}$ has dimension $n-k$. First, $\mathcal{C}$ can be written as the right kernel of a matrix $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$ with rank $n-k$,

$$\mathcal{C} = \left\{\mathbf{c} \in \mathbb{F}_q^n : \ \mathbf{H}\mathbf{c}^\top = \mathbf{0}\right\}$$

Therefore, all rows of $\mathbf{H}$ are elements in $\mathcal{C}^\perp$ showing that $\dim \mathcal{C}^\perp \geq n-k$. On the other hand, if $\mathbf{B} \in \mathbb{F}_q^{m\times n}$ is a basis $\big($considering its rows$\big)$ of $\mathcal{C}^\perp$. Then by linearity $\mathcal{C}$ is included in the $\big($right$\big)$ kernel of $\mathbf{B}$. We deduce that $k = \dim \mathcal{C} \leq n - \dim \mathcal{C}^\perp$ concluding the whole proof

$\mathsf{G} \in \mathbb{F}_q^{k \times n}$ generator matrix of $\mathcal{C}$, *i.e.*, $\mathcal{C} = \left\{\mathsf{mG} : \ \mathsf{m} \in \mathbb{F}_q^k\right\}$

$\longrightarrow$ SG is still a generator matrix when $\mathsf{S} \in \mathbb{F}_q^{k \times k}$ is invertible

$\mathsf{H} \in \mathbb{F}_q^{(n-k) \times n}$ parity-check matrix of $\mathcal{C}$, *i.e.*, $\mathcal{C} = \left\{\mathsf{c} \in \mathbb{F}_q^n : \ \mathsf{Hc}^\top = \mathsf{0}\right\}$

$\longrightarrow$ SH is still a parity-check matrix when $\mathsf{S} \in \mathbb{F}_q^{(n-k) \times (n-k)}$ is invertible

Left multiplication by an invertible matrix computes a change of basis!

$\mathbf{G} \in \mathbb{F}_q^{k \times n}$ generator matrix of $\mathcal{C}$ $\overset{\text{easy to compute?}}{\longleftrightarrow}$ $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$ parity-check matrix of $\mathcal{C}$

- Given $\mathbf{G} \in \mathbb{F}_q^{k \times n}$, it has rank k. We can perform a Gaussian elimination, *i.e.*, compute $\mathbf{S} \in \mathbb{F}_q^{k \times k}$ invertible such that $\Big($up to a permutation of columns$\Big)$,

$$\mathbf{SG} = (\mathbf{I}_k \mid \mathbf{A}) \text{ where } \mathbf{A} \in \mathbb{F}_q^{k \times (n-k)}$$

$$\longrightarrow \text{Then } \mathbf{H} = (-\mathbf{A}^\top \mid \mathbf{I}_{n-k}) \text{ parity-check matrix of } \mathcal{C}$$

Proof:

Indeed, $\mathbf{m}\,(\mathbf{SG})\,\mathbf{H}^\top = \mathbf{m}(\mathbf{I}_k \mid \mathbf{A}) \left(\frac{-\mathbf{A}}{\mathbf{I}_{n-k}}\right) = \mathbf{m}(\mathbf{0}) = (\mathbf{0})$. Therefore, $\mathcal{C}$ included in the right kernel of $\mathbf{H}$. But $\mathbf{H}$ has rank $n-k$, showing the $\mathbf{H}$ is a parity-check matrix of $\mathcal{C}$

$\mathbf{G} \in \mathbb{F}_q^{k \times n}$ generator matrix of $\mathcal{C}$ $\overset{\text{easy to compute?}}{\longleftrightarrow}$ $\mathbf{H} \in \mathbb{F}_q^{(n-k) \times n}$ parity-check matrix of $\mathcal{C}$

- Given $\mathbf{H} \in \mathbb{F}_q^{(n-k) \times n}$, it has rank $n-k$. We can perform a Gaussian elimination, *i.e.*, compute $\mathbf{S} \in \mathbb{F}_q^{(n-k) \times (n-k)}$ invertible such that $\Big($up to a permutation of columns$\Big)$,

$$\mathbf{S}\mathbf{H} = (\mathbf{I}_{n-k} \mid \mathbf{B}) \text{ where } \mathbf{B} \in \mathbb{F}_q^{(n-k) \times k}$$

$$\longrightarrow \text{Then } \mathbf{G} = (-\mathbf{B}^\top \mid \mathbf{I}_k) \text{ generator matrix of } \mathcal{C}$$

Exercise:

Given $\mathbf{x} \in \mathbb{F}_q^n$ and a linear code $\mathcal{C} \subseteq \mathbb{F}_q^n$, is it easy to decide if $\mathbf{x} \in \mathcal{C}$?

$$\mathcal{C}^\perp = \left\{ \mathbf{c}^\perp \in \mathcal{C}^\perp :\ \forall \mathbf{c} \in \mathcal{C},\ \langle \mathbf{c}, \mathbf{c}^\perp \rangle = \sum_{i=1}^{n} c_i^\perp c_i = 0 \in \mathbb{F}_q \right\}$$

If $\mathcal{C} \subseteq \mathbb{F}_q^n$ has dimension k, then $\mathcal{C}^\perp$ has dimension $n - k$ where $n = \mathbf{dim}\, \mathbb{F}_q^n$

$\longrightarrow$ It seems that $\mathcal{C}^\perp$ is the orthogonal of $\mathcal{C}$ and $\langle \cdot, \cdot \rangle$ is a scalar product, **but no!**

The dual is not an orthogonal!

$$\mathcal{C} + \mathcal{C}^\perp \neq \mathbb{F}_q^n$$

It happens that $\mathcal{C} \bigcap \mathcal{C}^\perp \neq \{0\}$ and this intersection is called the hull

Characters and Fourier transforms for prime q:

- Characters are $\chi_{\mathbf{x}}(\mathbf{y}) = e^{2i\pi \langle \mathbf{x},\mathbf{y}\rangle / q}$ where $\mathbf{x}, \mathbf{y} \in \mathbb{F}_q^n$. They are morphisms from $(\mathbb{F}_q^n, +)$ to the units of $(\mathbb{C}, \times)$.
- The Fourier transform of $f : \mathbb{F}_q^n \to \mathbb{C}$, is

$$\widehat{f}(\mathbf{x}) = \frac{1}{\sqrt{q}} \sum_{\mathbf{y}\in\mathbb{F}_q^n} f(y)\chi_{\mathbf{x}}(\mathbf{y})$$

The dual code is defined via group theory involving dual groups

The dual code $\mathcal{C}^\perp$ is the set of points for which characters are trivial when restricted to $\mathcal{C}$, *i.e.*,

$$\mathcal{C}^\perp = \left\{ \mathbf{c}^\perp \in \mathbb{F}_q^n : \ \forall \mathbf{c} \in \mathcal{C}, \ \chi_{\mathbf{c}^\perp}(\mathbf{c}) = 1 \right\}$$

$$\left(\text{when } q \text{ prime, } \chi_{\mathbf{x}}(\mathbf{y}) = 1 \iff \langle \mathbf{x}, \mathbf{y} \rangle = 0 \in \mathbb{F}_q \right)$$

Given two finite subspaces: $F \subseteq E$

Equivalence relation: $x \sim y \iff x - y \in F$

$$E/F = \{\overline{x} \ : \ x \in E\} \quad \text{where} \quad \overline{x} \overset{\text{def}}{=} \{y \in E \ : \ x \sim y\} = x + F$$

$\longrightarrow$ It defines a linear space!

$$k = \mathbf{dim}\, E/F = \mathbf{dim}\, E - \mathbf{dim}\, F$$

Rough analogy:

E/F	$\mathbb{Z}/4\mathbb{Z}$
$\{\overline{x_1}, \dots, \overline{x_N}\}$	$\{\overline{0}, \overline{1}, \overline{2}, \overline{3}\}$
$\overline{x_i} = x_i + F$	$\overline{\ell} = \ell + 4\mathbb{Z}$
$\overline{x} = \overline{y} \iff x - y \in F$	$\overline{\ell} = \overline{m} \iff \ell - m \in 4\mathbb{Z}$
$E = \bigsqcup_{1 \leq i \leq N} \overline{x_i}$	$\mathbb{Z} = \bigsqcup_{\ell \in \{0,1,2,3\}} \overline{\ell}$

Decoding: given $\mathbf{c} + \mathbf{e}$, recover $\mathbf{e}$

$\longrightarrow$ Make **modulo $\mathcal{C}$** to extract the information about $\mathbf{e}$

Coset space: $\mathbb{F}_2^n/\mathcal{C}$

$$\sharp\, \mathbb{F}_q^n/\mathcal{C} = q^{n-k} \quad \text{and} \quad \mathbb{F}_q^n/\mathcal{C} = \left\{ \overline{\mathbf{x}}_i \;: 1 \leq i \leq q^{n-k} \right\} = \left\{ \mathbf{x}_i + \mathcal{C} \;:\; 1 \leq i \leq q^{n-k} \right\}$$

where the $\mathbf{x}_i$'s are the **representatives** of $\mathbb{F}_q^n/\mathcal{C}$. The $x_i + \mathcal{C}$'s **are disjoint**!

A natural set of representatives via a parity-check $\mathbf{H}$: **syndromes**

Proposition:

We have:

1. $\mathbf{x}_i + \mathcal{C} \in \mathbb{F}_q^n/\mathcal{C} \longmapsto \mathbf{H}\mathbf{x}_i^{\mathsf{T}} \in \mathbb{F}_q^{n-k}$ $\Big($called a syndrome$\Big)$ is an isomorphism
2. $\mathbb{F}_q^n = \bigsqcup_{\mathbf{s} \in \mathbb{F}_q^{n-k}} \left\{ \mathbf{z} \in \mathbb{F}_q^n \;:\; \mathbf{H}\mathbf{z}^{\mathsf{T}} = \mathbf{s}^{\mathsf{T}} \right\}$

$$\mathbf{c} + \mathbf{e} \bmod \mathcal{C} = \mathbf{H}(\mathbf{c} + \mathbf{e})^{\mathsf{T}} = \underbrace{\mathbf{H}\mathbf{c}^{\mathsf{T}}}_{=0} + \mathbf{H}\mathbf{e}^{\mathsf{T}} = \mathbf{H}\mathbf{e}^{\mathsf{T}}$$ which gives information to recover $\mathbf{e}$ $\Big($decoding$\Big)$

$\longrightarrow$ $\mathbf{c} + \mathbf{e} \bmod \mathcal{C}$ is only function of $\mathbf{e}$!

Proof:

1. Let us first show that $\mathbf{x}_i + \mathcal{C} \in \mathbb{F}_q^n/\mathcal{C} \longmapsto \mathbf{H}\mathbf{x}_i^{\mathsf{T}} \in \mathbb{F}_q^{n-k}$ is a well-defined mapping. If we choose another class representative $\mathbf{y}_i + \mathcal{C} = \mathbf{x}_i + \mathcal{C}$. Then by definition

$$\mathbf{y}_i - \mathbf{x}_i \in \mathcal{C} \iff \mathbf{H}\left(\mathbf{y}_i - \mathbf{x}_i\right)^\top = 0 \iff \mathbf{H}\mathbf{y}_i^\top = \mathbf{H}\mathbf{x}_i^\top$$

   It shows that we have a well-defined mapping. But the equivalence also shows that it is a one-to-one mapping
   The above application is surjective as $\mathbf{H}$ has rank $n-k$, therefore for any $\mathbf{s} \in \mathbb{F}_q^{n-k}$ it exists $\mathbf{x} \in \mathbb{F}_q^n$ such that $\mathbf{H}\mathbf{x}^\top = \mathbf{s}^\top$ and $\mathbf{x}$ defines one representative. Furthermore the mapping is clearly linear, concluding the proof of 1

2. This is a consequence of the equivalence relation but let's give a direct proof. We have shown above that $\forall \mathbf{z} \in \mathbb{F}_q^n$, it exists $\mathbf{s} \in \mathbb{F}_q^n$ such that $\mathbf{H}\mathbf{z}^\top = \mathbf{s}^\top$ $\Big(\mathbf{H}$ has rank $n-k\Big)$.
   To conclude notice that $\left\{\mathbf{z} \in \mathbb{F}_q^n \ : \ \mathbf{H}\mathbf{z}^{\mathsf{T}} = \mathbf{s}^{\mathsf{T}}\right\}$ are clearly disjoint for $\mathbf{s} \in \mathbb{F}_q^{n-k}$

$\mathcal{C}$ be an $[n, k]_q$-code with generator and parity-check matrices $\mathbf{G}$ and $\mathbf{H}$

- Given a noisy codeword, $\mathbf{y} = \underbrace{\mathbf{c}}_{\in \mathcal{C}} + \mathbf{e}$, its syndrome is

$$\mathbf{H}\mathbf{y}^{\top} = \mathbf{H}\mathbf{c}^{\top} + \mathbf{H}\mathbf{e}^{\top} = \mathbf{H}\mathbf{e}^{\top} \text{ where we use } \mathcal{C} = \left\{ \mathbf{c} \in \mathbb{F}_q^n : \ \mathbf{H}\mathbf{c}^{\top} = \mathbf{0} \right\}$$

- Given a syndrome, $\mathbf{s}^{\top} = \mathbf{H}\mathbf{e}^{\top}$, we can easily compute its associated noisy codeword, by a Gaussian elimination we compute $\mathbf{y}$ such that $\mathbf{H}\mathbf{y}^{\top} = \mathbf{s}^{\top}$ $\left(\text{as } \mathbf{rank}(\mathbf{H}) = n - k\right)$

$$\mathbf{H}\mathbf{y}^{\top} = \mathbf{s}^{\top} \iff \mathbf{H}(\mathbf{y} - \mathbf{e})^{\top} = \mathbf{0} \iff \mathbf{y} - \mathbf{e} \in \mathcal{C} \iff \mathbf{y} = \underbrace{\mathbf{c}}_{\in \mathcal{C}} + \mathbf{e}$$

HAMMING DISTANCE

OUR GOAL

Remember, we introduced codes to communicate over a noisy channel

$\longrightarrow$ We will restrict our attention to the following channels:

q-ary symmetric channels:

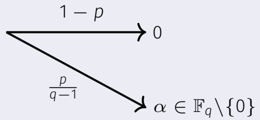
Memoryless channel $\left(\mathbb{F}_q, \mathbb{F}_q, p(y \mid x)\right)$ where,

$$\forall x, y \in \mathbb{F}_q, \quad p(y \mid x) = \begin{cases} 1 - p \text{ if } x = y \\ \frac{p}{q-1} \text{ otherwise} \end{cases}$$

p: probability of error ; $\frac{p}{q-1}$ transition probability

When sending $\mathbf{x} \in \mathbb{F}_q^n$ through the channel

$$\mathbf{y} = \mathbf{c} + \mathbf{e} \quad \text{where the } e_i \text{ are i.i.d and } p(e_i = x) = \begin{cases} 1 - p \text{ if } x = 0 \\ \frac{p}{q-1} \text{ otherwise} \end{cases}$$



Remember, after sending a codeword across a noisy channel we want to recover the sent codeword, i.e., decoding

$\longrightarrow$ The optimal decoder is the following one:

Maximum likelihood decoder:

Given a q-ary symmetric channel $\left(\mathbb{F}_q, \mathbb{F}_q, p(y \mid x)\right)$ and a block-code $\mathcal{C} \subseteq \mathbb{F}_q^n$. We call the maximum likelihood decoder the map

$$\varphi : \mathbb{F}_q^n \longrightarrow \mathcal{C}$$

such that given $\mathbf{y} \in \mathbb{F}_q^n$, it outputs the codeword $\mathbf{c} \in \mathcal{C}$ maximizing the transition probabilities

$$\varphi(\mathbf{y}) \stackrel{\text{def}}{=} \underset{\mathbf{c} \in \mathcal{C}}{\arg\max}\, p(\mathbf{c} \mid \mathbf{y}) = \underset{\mathbf{c} \in \mathcal{C}}{\arg\max} \prod_{i=1}^{n} p(c_i \mid y_i)$$

Proposition:

In a q-ary symmetric channel with probability of transition $p/(q-1) < 1/q$, if codewords $\mathbf{c} \in \mathcal{C}$ are chosen uniformly at random among $\mathcal{C}$, then

$$\forall \mathbf{y} \in \mathbb{F}_q^n,\ \varphi(\mathbf{y}) = \mathbf{c} \in \mathcal{C} \ \text{ such that } \mathbf{c} = \operatorname*{arg\,min}_{\mathbf{d} \in \mathcal{C}} d_{\mathrm{H}}(\mathbf{y}, \mathbf{d})$$

where $d_{\mathrm{H}}(\cdot,\cdot)$ is the Hamming distance,

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{F}_q^n,\quad d_{\mathrm{H}}(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \sharp \{ i \in [1, n],\ x_i \neq y_i \}$$

Given $\mathbf{y} \in \mathbb{F}_q^n$, the maximum likelihood decoder for q-ary symmetric channels outputs the closest codewords $\mathbf{c}$ for the Hamming distance

It justifies the use of the Hamming distance for decoding in the q-ary symmetric channel

Proof:

1. First, by using that the $\mathbf{c}$'s are uniform among $\mathcal{C}$

$$p(\mathbf{c} \mid \mathbf{y}) = p(\mathbf{y} \mid \mathbf{c}) \, \frac{p(\mathbf{c})}{p(\mathbf{y})} = p(\mathbf{y} \mid \mathbf{c}) \, \frac{1}{\sharp\mathcal{C} \; p(\mathbf{y})}$$

   We deduce that maximizing $p(\mathbf{c} \mid \mathbf{y})$ $\left(\text{over } \mathbf{c}\right)$ boils down to maximize $p(\mathbf{y} \mid \mathbf{c})$

2. Second, by definition of the q-ary symmetric channel,

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{c}) &= (1-p)^{\sharp\{i \in [1,n]:\ y_i = c_i\}} \left(\frac{p}{q-1}\right)^{\sharp\{i \in [1,n]:\ y_i \neq c_i\}} \\ &= (1-p)^{n - d_{\mathrm{H}}(\mathbf{y},\mathbf{c})} \left(\frac{p}{q-1}\right)^{d_{\mathrm{H}}(\mathbf{y},\mathbf{c})} \end{aligned}$$

   As $p/(q-1) < 1/q$,

$$\alpha \mapsto (1-p)^{n-\alpha} \left(\tfrac{p}{q-1}\right)^{\alpha}$$

   is a decreasing function showing that $p(\mathbf{y} \mid \mathbf{c})$ is maximal when $d_{\mathrm{H}}(\mathbf{y}, \mathbf{c})$ is minimal

Hamming weight:

$$\forall \mathbf{x} \in \mathbb{F}_q^n, \quad |\mathbf{x}| \stackrel{\text{def}}{=} \sharp \{i \in [1, n], \ x_i \neq 0\}$$

$$\longrightarrow \quad d_{\mathrm{H}}(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|$$

Some remarks:

- $|\cdot|$ is not a norm but $d_{\mathrm{H}}(\cdot, \cdot)$ is a distance
- The Hamming weight does not discriminate non-zero symbols, for instance in $\mathbb{F}_5 = \mathbb{Z}/5\mathbb{Z}$,

$$\Big|(1, 2, 0, 1, 0, 0, 2)\Big| = \Big|(3, 3, 4, 0, 0, 0, 1)\Big| = 4$$

An important parameter for a code: its minimum distance

$\longrightarrow$ It measures the quality of a code in terms of "error detection"

Minimum distance:

Given $\mathcal{C} \subseteq \mathbb{F}_q^n$, its minimum distance is defined as

$$d_{\min}(\mathcal{C}) \stackrel{\text{def}}{=} \min \left\{ |\mathbf{c}_1 - \mathbf{c}_2| : \mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C} \text{ and } \mathbf{c}_1 \neq \mathbf{c}_2 \right\}$$

Remark:

For a linear code $\mathcal{C}$,

$$d_{\min}(\mathcal{C}) = \min \left\{ |\mathbf{c}| : \mathbf{c} \in \mathcal{C} \backslash \{\mathbf{0}\} \right\}$$

Suppose that someone sends us a codeword $\mathbf{c} \in \mathcal{C}$ across a noisy channel

Our goal is to guess if an error occurred

How can we proceed? What is the maximal amount of errors for which we can take the right decision with certainty?

Suppose that someone sends us a codeword $\mathbf{c} \in \mathcal{C}$ across a noisy channel

Our goal is to guess if an error occurred

How can we proceed? What is the maximal amount of errors for which we can take the right decision with certainty?

Error detection strategy:

Given a received $\mathbf{y}$ we compute $\mathbf{H}\mathbf{y}^\top$ for $\mathbf{H}$ being a parity-check matrix of the code. If we obtain $\mathbf{0}$ then we say that no error occurred

This strategy gives the right answer with certainty if the Hamming weight of the error is $< d_{\min}(\mathcal{C})$!

Proof:

If an error occurred then we receive $\mathbf{c} + \mathbf{e}$. Therefore $\mathbf{H}\left(\mathbf{c} + \mathbf{e}\right)^\top = \mathbf{H}\mathbf{c}^\top + \mathbf{H}\mathbf{e}^\top = \mathbf{H}\mathbf{e}^\top$. Then if $|\mathbf{e}| < d_{\min}(\mathcal{C})$ we necessarily have $\mathbf{e} \notin \mathcal{C}$ and $\mathbf{H}\mathbf{e}^\top \neq \mathbf{0}$. However, if $|\mathbf{e}| \geq d_{\min}(\mathcal{C})$ it is possible that $\mathbf{e} \in \mathcal{C}$ and $\mathbf{H}\mathbf{e}^\top = \mathbf{0}$

$$\mathbf{x} \in \mathbb{F}_q^n, \quad \mathcal{B}(\mathbf{x}, r) \stackrel{\text{def}}{=} \left\{\mathbf{y} \in \mathbb{F}_q^n : \ |\mathbf{y} - \mathbf{x}| \leq r\right\}$$

Proposition:

Given a code $\mathcal{C} \subseteq \mathbb{F}_q^n$,

$$\forall \mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}, \quad \mathbf{c}_1 \neq \mathbf{c}_2: \qquad \mathcal{B}\left(\mathbf{c}_1, \left\lfloor \frac{d_{\min}(\mathcal{C})-1}{2} \right\rfloor\right) \bigcap \mathcal{B}\left(\mathbf{c}_2, \left\lfloor \frac{d_{\min}(\mathcal{C})-1}{2} \right\rfloor\right) = \emptyset$$

Proof:

By contradiction, suppose there exists $\mathbf{y} \in \mathcal{B}\left(\mathbf{c}_1, \left\lfloor \frac{d_{\min}(\mathcal{C})-1}{2} \right\rfloor\right) \bigcap \mathcal{B}\left(\mathbf{c}_2, \left\lfloor \frac{d_{\min}(\mathcal{C})-1}{2} \right\rfloor\right)$,

$$\begin{aligned} |\mathbf{c}_1 - \mathbf{c}_2| &= |(\mathbf{c}_1 - \mathbf{y}) - (\mathbf{c}_2 - \mathbf{y})| \\ &\leq |\mathbf{c}_1 - \mathbf{y}| + |\mathbf{c}_2 - \mathbf{y}| \qquad \text{(triangular inequality)} \\ &\leq \left\lfloor \frac{d_{\min}(\mathcal{C}) - 1}{2} \right\rfloor + \left\lfloor \frac{d_{\min}(\mathcal{C}) - 1}{2} \right\rfloor \\ &< d_{\min}(\mathcal{C}) \end{aligned}$$

which is a contradiction as $\mathbf{c}_1 \neq \mathbf{c}_2$ and they belong to $\mathcal{C}$ with minimum distance $d_{\min}(\mathcal{C})$

When transmitting $\mathbf{c} \in \mathcal{C}$, if the Hamming weight of the error is $< d_{\min}(\mathcal{C})/2$, then the maximum likelihood decoder necessarily outputs $\mathbf{c}$ $\Big($closest codeword for the Hamming distance$\Big)$

When transmitting $\mathbf{c} \in \mathcal{C}$, if the Hamming weight of the error is $< d_{\min}(\mathcal{C})/2$, then the maximum likelihood decoder necessarily outputs $\mathbf{c}$

> The above statement says that with $< d_{\min}(\mathcal{C})/2$ error the maximum likelihood decoder succeeds **with certainty**!

$\longrightarrow$ There are codes for which the maximum likelihood decoder works with probability $1 - \mathrm{e}^{-Cn}$ as soon as there are $\leq d_{\min}(\mathcal{C})$ errors, we gain a factor two!

$\Big($in particular random codes as we will see in Lecture 8$\Big)$

The minimum distance quantifies how "good" is a code in term of error decoding/detection

- Balls centered at codewords with radius $\left\lfloor \frac{d_{\min}(\mathcal{C})-1}{2} \right\rfloor$ are disjoint

  $\longrightarrow$ We can **correct** $\left\lfloor \frac{d_{\min}(\mathcal{C})-1}{2} \right\rfloor$ errors **with certainty**!

- There are no codewords in any ball centered at codewords with radius $d_{\min}(\mathcal{C}) - 1$

  $\longrightarrow$ We can **detect** any $< d_{\min}(\mathcal{C})$ errors

Given an $[n, k]_q$-code $\mathcal{C}$, how large can be its minimum distance $d_{\min}(\mathcal{C})$?

BOUNDS ON MINIMUM DISTANCE

The $[7,4]_2$ Hamming code $\mathcal{C}_\mathrm{H}$ admits as parity check matrix

$$\mathbf{H} \stackrel{\text{def}}{=} \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

It has minimum distance $d_{\mathrm{min}}(\mathcal{C}_\mathrm{H}) = 2$, indeed use the following proposition

Proposition:

Given a linear code $\mathcal{C}$ with parity-check matrix $\mathbf{H}$,

$$\mathcal{C} \text{ has minimum distance } \geq d \iff \text{every } d-1 \text{ columns of } \mathbf{H} \text{ form a free family}$$

$$\text{Given } \mathbf{c} \in \mathcal{C}_\mathrm{H} \text{ there are } 8 = 2^3 \text{ noisy codewords } \mathbf{c}+\mathbf{e} \text{ where } |\mathbf{e}| \leq 1 = \left\lfloor \frac{d_{\mathrm{min}}(\mathcal{C}_\mathrm{H})-1}{2} \right\rfloor$$

$$\longrightarrow \text{The balls } \mathcal{B}\left(\mathbf{c}, \left\lfloor \frac{d_{\mathrm{min}}(\mathcal{C})-1}{2} \right\rfloor\right)\text{'s for } \mathbf{c} \in \mathcal{C}_\mathrm{H} \text{ form a partition of } \mathbb{F}_2^7!$$

Perfect codes:

A linear code $\mathcal{C} \subseteq \mathbb{F}_q^n$ is said to be perfect if the balls $\mathcal{B}\left(\mathbf{c}, \left\lfloor \frac{d_{\mathrm{min}}(\mathcal{C})-1}{2} \right\rfloor\right)$ form a partition of $\mathbb{F}_q^n$

Exercise:

Given a parity-check of some matrix $\mathbf{H}$, is it easy to check that every $d-1$ columns of $\mathbf{H}$ form a free family? More generally, does it seem easy to compute the minimum distance of a given code?

An Hamming code is the $[2^m - 1, 2^m - m - 1]_2$-code admitting as parity-check matrix $\mathbf{H} \in \mathbb{F}_2^{(2^m-1)\times m}$ whose columns are all the vectors $\mathbb{F}_2^m \backslash \{\mathbf{0}\}$. It has minimum distance 3 and it is a perfect code

$$2^{2^m - m - 1}\left(\binom{2^m-1}{1} + 1\right) = 2^{2^m - 1}$$

Theorem:

Parameters $[n, k, d_{\min}(\mathcal{C})]_q$ of perfect codes are known: $[2\ell + 1, 1, 2\ell + 1]_2$ $\big($repetition codes with odd length$\big)$, $[2^m - 1, 2^m - m - 1, 3]_2$ $\big($Hamming codes$\big)$, $[23, 12, 7]_2$ $\big($binary Golay code $G_{23}\big)$ and $[11, 6, 5]_3$ $\big($ternary Golay code $G_{11}\big)$

Singleton bound:

For all $[n, k]_q$-code $\mathcal{C}$,

$$d_{\min}(\mathcal{C}) \leq n - k + 1$$

Proof:

Given a parity-check matrix $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$, it has rank $n - k$. We cannot hope having more than $(n - k)$ columns forming a free family. Therefore,

$$d - 1 \leq n - k$$

Do we know codes that reach this bound? Yes!

(codes reaching the Singleton bound are said MDS, *i.e.*, Maximum Distance Separable)

Reed-Solomon code:

Given $x_1, \ldots, x_n \in \mathbb{F}_q$ where the x_i's are different, *i.e.*, $x_i \neq x_j$,

$$\left\{ (f(x_1), \ldots, f(x_n)) : \; f \in \mathbb{F}_q[X], \; \deg(f) < k \right\}$$

is a $[n, k]_q$-code with minimum distance $n - k + 1$

Reed-Solomon codes have optimal minimum distances, **but be careful, their length $n \leq q$**

$\longrightarrow$ There are sharper bounds when q is fixed and n grows!

Hamming bound:

For any code $\mathcal{C} \subseteq \mathbb{F}_q^n$,

$$\sharp\mathcal{C} \cdot \sharp\mathcal{B}\left(0, \left\lfloor \frac{d_{\min}(\mathcal{C})-1}{2} \right\rfloor\right) = \sharp\mathcal{C} \cdot \left(\sum_{r=0}^{\left\lfloor \frac{d_{\min}(\mathcal{C})-1}{2} \right\rfloor} \binom{n}{r}(q-1)^r \right) \leq q^n$$

It asymptotic form when $n \to +\infty$: for any sequence of codes $\mathcal{C}_n \subseteq \mathbb{F}_q^n$ such that the following limits exist:

$$\delta \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{d_{\min}(\mathcal{C})}{n} \quad \text{and} \quad R \stackrel{\text{def}}{=} \lim_{n\to+\infty} \frac{\log_q \sharp\mathcal{C}_n}{n}$$

we have,

$$\frac{\delta}{2} \leq h_q^{-1}(1-R) \quad \text{where } h_q(x) \stackrel{\text{def}}{=} -(1-x)\log_q(1-x) - x\log_q \frac{x}{q-1}$$

Proof:

The Hamming bound is a consequence of the fact that balls centered at codewords with radius $\left\lfloor \frac{d_{\min}(\mathcal{C})-1}{2} \right\rfloor$ are disjoint

The asymptotic form comes from the fact that for a fixed q, $\binom{n}{r}(q-1)^n = \text{poly}(n) \cdot q^{nh_q(r/n)}$

Gilbert-Varshamov bound:

$q^k \cdot \sharp\mathcal{B}(0, d-2) = q^k \cdot \sum_{i=0}^{d-2} \binom{n}{i}(q-1)^i < q^n \implies$ it exits an $[n, k]_q$-code with minimum distance d

$\longrightarrow$ The maximum d reaching the inequality is $d_{\mathrm{GV}}(n, k)$

Be careful:

The Gilbert-Varshamov bound states that it exists an $[n, k]_q$-code $\mathcal{C}$ with $d_{\min}(\mathcal{C}) \geq d_{\mathrm{GV}}(n, k)$, **not that for all $[n, k]_q$-code $\mathcal{C}$, $d_{\min}(\mathcal{C}) \leq d_{\mathrm{GV}}(n, k)$**

*Almost all codes reach **asymptotically** the Gilbert-Varshamov bound*

Asymptotic Gilbert-Varshamov bound:

Let $\varepsilon > 0$ and $\delta_{\mathrm{GV}} = h_q^{-1}(1-R)$. We have for uniform $[n, Rn]_q$-codes $\mathcal{C}$,

$$\mathbb{P}_{\mathcal{C}}\left((1-\varepsilon)\delta_{\mathrm{GV}} < \frac{d_{\min}(\mathcal{C})}{n} < (1+\varepsilon)\delta_{\mathrm{GV}}\right) \geq 1 - q^{-\alpha n(1+o(1))}$$

where $\alpha \stackrel{\text{def}}{=} \min\left((1-R) - h_q\left((1+\varepsilon)\delta_{\mathrm{GV}}\right), h_q\left((1-\varepsilon)\delta_{\mathrm{GV}}\right) - (1-R)\right) > 0$

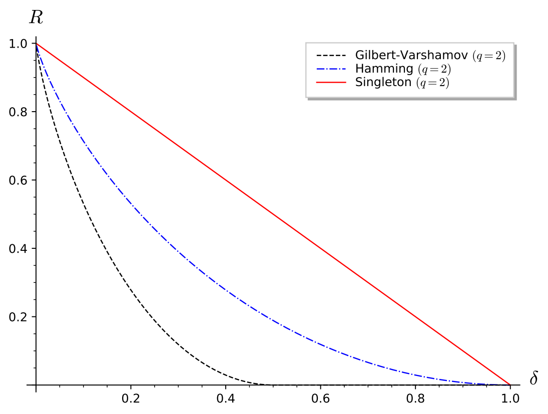
1M$ open question:

Does it exist a sequence of binary linear codes $\mathcal{C}_n$ with rate R such that

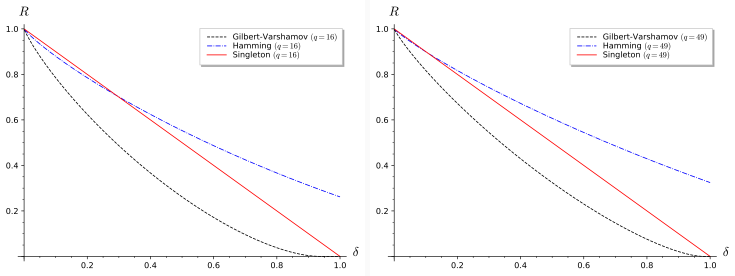
$$\frac{d_{\min}(\mathcal{C}_n)}{n} \xrightarrow[n \to +\infty]{} \delta > \delta_{\mathrm{GV}} = h_2^{-1}(1-R)?$$

Bounds for sequences of $[n, Rn]_q$-codes $\mathcal{C}_n$ s.t. $\frac{d_{\mathsf{min}}(\mathcal{C}_n)}{n} \xrightarrow[n\to+\infty]{} \delta$ $\big($but q is fixed$\big)$

$\big($It exists codes above Gilbert-Varshamov, all codes are below Hamming and Singleton$\big)$



Bounds for sequences of $[n, Rn]_q$-codes $\mathcal{C}_n$ s.t. $\frac{d_{\min}(\mathcal{C}_n)}{n} \xrightarrow[n\to+\infty]{} \delta$ $\Big($but q is fixed$\Big)$



NOT A PARADOX

We have seen that Reed-Solomon codes reach the Singleton bound (*red curve*)

But the Hamming bound (*blue curve*) is an upper-bound below the Singleton bound

Don't forget that for Reed-Solomon codes we have $q \geq n$ and for our curve we let $n \to +\infty$ while q is fixed!

DECODING REED-SOLOMON CODES

Reed-Solomon (RS) codes:

$\mathbf{x} \in \mathbb{F}_q^n$ such that $x_i \neq x_j$ $\left(\text{in particular } n \leq q\right)$ and $k \leq n$. The code $\mathrm{RS}_k(\mathbf{x})$ is defined as

$$\mathrm{RS}_k(\mathbf{x}) \stackrel{\text{def}}{=} \left\{(f(x_1), \dots, f(x_n)) : \ f \in \mathbb{F}_q[X] \text{ and } \mathbf{deg}(f) < k\right\}$$

$\longrightarrow$ These codes are used in QR-codes!

Exercise:

Show that $\mathrm{RS}_k(\mathbf{x})$ has generator matrix

$$\mathbf{G} \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{k-1} & x_2^{k-1} & \cdots & x_n^{k-1} \end{pmatrix}$$

Decoding algorithm:

Given, $\mathrm{RS}_k(\mathbf{x})$ and $\mathbf{c}+\mathbf{e}$ such that $\begin{cases} \mathbf{c} \in \mathrm{RS}_k(\mathbf{x}) \\ |\mathbf{e}| \leq \left\lfloor \frac{n-k}{2} \right\rfloor \end{cases}$

Then, we can efficiently recover $(\mathbf{c}, \mathbf{e})$

$$\text{Given } \mathbf{y} = \mathbf{c} + \mathbf{e} \text{ where } \begin{cases} \mathbf{c} \in \mathsf{RS}_k(\mathbf{x}) \\ |\mathbf{e}| \leq \left\lfloor \frac{n-k}{2} \right\rfloor \end{cases}.$$

By definition, $\mathbf{c} = \left(f(x_i)\right)_i$ where $f \in \mathbb{F}_q[X]$ **is unknown** with $\mathbf{deg}(f) < k$

1. Let $\mathcal{I}$ be the **unknown** set of positions where $e_i \neq 0$, *i.e.*,

$$\mathcal{I} = \left\{ i \in [1,n] : \; y_i \neq f(x_i) \right\}$$

Fundamental idea (I):

Let $E \in \mathbb{F}_q[X]$ be the following **unknown** polynomial,

$$E(X) = \prod_{i \in \mathcal{I}} (X - x_i) \quad \text{which has degree} \leq \left\lfloor \frac{n-k}{2} \right\rfloor \text{ by assumption on } |\mathbf{e}|$$

2. By definition of $\mathcal{I}$ and E,

$$\forall i \in [1,n], \quad y_i E(x_i) = f(x_i) E(x_i)$$

3. The x_i's and y_i's are known: we have above a quadratic system to solve which is a priori not easy

2. By definition of $\mathcal{I}$ and E,

$$\forall i \in [1, n], \quad y_i E(x_i) = f(x_i) E(x_i)$$

3. The x_i's and y_i's are known, we have above a quadratic system to solve which is not easy

Fundamental idea (II): linearize

Solve the following linear system for **unknown** $N \in \mathbb{F}_q[X]$ with degree $\leq k - 1 + \left\lfloor \frac{n-k}{2} \right\rfloor$,

$$\forall i \in [1, n], \quad y_i E(x_i) = N(x_i)$$

There are n equations and $k + 2 \left\lfloor \frac{n-k}{2} \right\rfloor + 1$ **unknowns** $\left(\text{coefficients of } N \text{ and } E\right)$

$\longrightarrow (E, Ef)$ is a solution but it it is not the only one. . .

Fundamental idea (II): linearize

Solve the following linear system for **unknown** $N \in \mathbb{F}_q[X]$ with degree $\leq k - 1 + \left\lfloor \frac{n-k}{2} \right\rfloor$ and $E \in \mathbb{F}_q[X]$ with degree $\leq \left\lfloor \frac{n-k}{2} \right\rfloor$,

$$\forall i \in [1, n],\ y_i E(x_i) = N(x_i) \tag{1}$$

Lemma:

Any non-zero solution (E_1, N_1) and (E_2, N_2) of the above system is such that $\frac{N_1}{E_1} = \frac{N_2}{E_2} = f$

$\longrightarrow$ Therefore the decoding algorithm only consists in computing an non-zero solution (E, N) and to output $f = N/E$

Proof:

First, if $E_i = 0$, then by Equation (1), N_i has $n > k - 1 + \lfloor (n-k)/2 \rfloor$ zeros and $N_i = 0$. Therefore, $E_i \neq 0$. Now set $R = N_1 E_2 - N_2 E_1$. We have,

$$\mathbf{deg}(R) \leq k - 1 + 2 \left\lfloor \frac{n-k}{2} \right\rfloor \leq n - 1$$

On the other hand, by Equation (1),

$$\forall i \in [1, n],\ R(x_i) = N_1(x_i)E_2(x_i) - N_2(x_i)E_1(x_i) = y_i E_1(x_i)E_2(x_i) - y_i E_1(x_i)E_2(x_i) = 0$$

Therefore, $R = 0$, showing $N_1/E_1 = N_2/E_2$. But (N, Ef) is a non-zero solution concluding the proof

We have demonstrated that we can decode Reed-Solomon codes. Does it exist other codes that we know how to decode?

$\longrightarrow$ Yes!

- **Algebraic decoders:** Reed-Solomon, alternant, geometric codes
- **Probabilistic decoders:** LDPC, Turbo, Polar codes

Be careful:

- It is an hard problem to design codes with an efficient decoding algorithm
- When designing a decoding algorithm we have to be cautious about the parameters: the field size q or the decoding distance $\big($larger it is, harder is to decode$\big)$

FURTHER READING

- An introduction to Low-Density Parity-Check (LDPC) codes available here:

  https://repository.arizona.edu/handle/10150/607470

EXERCISE SESSION