Linear codes

CoCoNut, 2016
Emmanuela Orsini

Previously . . .

- Block codes
  - ◇ Parameters: length, information rate, minimum distance
  - ◇ Examples: Parity code, Hamming code
- (MLD) $max_{c\in C}\,\Pr(\mathbf{r}|\mathbf{c}) \cong \min_{c\in C} d(\mathbf{r},\mathbf{c})$ (MMD)
- Binary Symmetric Channel (BSC)

Binary Symmetric Channel

Suppose $\mathbf{c}$ is the transmitted codeword and $\mathbf{r}$ is the received word:

$$\mathbf{c} = \mathbf{r} + \mathbf{e}$$

Binary Symmetric Channel

Suppose $\mathbf{c}$ is the transmitted codeword and $\mathbf{r}$ is the received word:

$$\mathbf{c} = \mathbf{r} + \mathbf{e}$$

Given two codewords $\mathbf{c}_1, \mathbf{c}_2$, then

$$\begin{aligned} \Pr(\mathbf{r}|\mathbf{c}_1) \leq \Pr(\mathbf{r}|\mathbf{c}_2) &\iff d(\mathbf{r}, \mathbf{c}_1) \geq d(\mathbf{r}, \mathbf{c}_2) \\ &\iff \mathrm{wt}(\mathbf{r} + \mathbf{c}_1) \geq \mathrm{wt}(\mathbf{r} + \mathbf{c}_2) \\ &\iff \mathrm{wt}(\mathbf{e}_1) \geq \mathrm{wt}(\mathbf{e}_2) \end{aligned}$$

The most likely codeword sent is the one corresponding to the error of smallest weight

Do we need more structure?

Binary Hamming code $(7,16)$: $\mathrm{Enc} : \{0,1\}^4 \to \{0,1\}^7$

Information bits	Codeword	Information bits	Codeword
0000	0000000	1000	1000110
0001	0001111	1001	1001001
0100	0010101	1010	1010101
0011	0011100	1011	1011010
0010	0010011	1100	1100011
0101	0101010	1101	1101100
0110	0110110	1110	1110000
0111	0111001	1111	1111111

We need $n \cdot 2^k$ bits to store a binary code $\mathrm{Enc} : \{0,1\}^k \to \{0,1\}^n$

Can we do better than this?

Do we need more structure?

Binary Hamming code $(7, 16)$: $\mathrm{Enc} : \{0,1\}^4 \to \{0,1\}^7$

Information bits	Codeword	Information bits	Codeword
0000	0000000	1000	1000110
0001	0001111	1001	1001001
0100	0010101	1010	1010101
0011	0011100	1011	1011010
0010	0010011	1100	1100011
0101	0101010	1101	1101100
0110	0110110	1110	1110000
0111	0111001	1111	1111111

We need $n \cdot 2^k$ bits to store a binary code $\mathrm{Enc} : \{0,1\}^k \to \{0,1\}^n$

Can we do better than this?

We need extra structure that would facilitate a succinct representation of the code

Can we do better?

Mathematically we can describe the $(7, 16)_2$ Hamming code by a matrix

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

so that, if we represent a message by the vector $\mathbf{m} = (m_1\ m_2\ m_3\ m_4)$, we can encode by computing

$$\mathbf{c} = \mathbf{m} \cdot G$$

Suppose we wish to transmit $\mathbf{m} = (1\ 0\ 1\ 0)$, we then compute

$$(1\,0\,1\,0) \cdot \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} = (1\,0\,1\,0\,1\,0\,1)$$

Can we do better?

$$(1\,0\,1\,0) \cdot \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} = (1\,0\,1\,0\,1\,0\,1)$$

Information bits	Codeword	Information bits	Codeword
0000	0000000	1000	1000110
0001	0001111	1001	1001001
0100	0010101	**1010**	**1010101**
0011	0011100	1011	1011010
0010	0010011	1100	1100011
0101	0101010	1101	1101100
0110	0110110	1110	1110000
0111	0111001	1111	1111111

Linear codes - Definition

The previous example is an example of **linear code**.

Definition (Linear code)

Let q be a prime power. Then $C \subseteq \{0, 1, \dots, q-1\}^n = \mathbb{F}_q^n$ is a linear code if it is a linear subspace of $\mathbb{F}_q^n$. If C has dimension k and distance d then it will be referred to as an $[n, k, d]_q$ or just an $[n, k]_q$ code.

- $\mathbb{F}_q^n$ denote the vector space of all n-tuples over the finite field $\mathbb{F}_q$.

Representing linear code

An $[n, k, d]_q$ code C is a subspace of $\mathbb{F}_q^n$.
We have two alternate characterization of C.

1. C is generated by its $k \times n$ **generator matrix** G, i.e. a matrix whose k rows span C.
   - The encoding map $\mathrm{Enc} : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is an injective linear map defined as

$$\mathbf{m} \mapsto \mathbf{m}G (= \mathbf{c})$$

Representing linear code

An $[n, k, d]_q$ code C is a subspace of $\mathbb{F}_q^n$.
We have two alternate characterization of C.

1. C is generated by its $k \times n$ **generator matrix** G, i.e. a matrix whose k rows span C.
   - The encoding map $\mathrm{Enc} : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is an injective linear map defined as
   $$\mathbf{m} \mapsto \mathbf{m}G (= \mathbf{c})$$
2. C is characterized by an $(n-k) \times n$ **parity-check matrix** H:
   $$C = \{\mathbf{c} \in \mathbb{F}_q^n \mid H\mathbf{c}^T = 0\}$$

Fact

The generator matrix and the parity-check matrix are orthogonal, i.e. $G \cdot H^T = 0$

Representing linear code - An example

The $[7,4,3]_2$ Hamming code has the following generator matrix

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

and the following parity-check matrix

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

- Both the generator matrix and the parity-check matrix can be represented using $O(n^2)$ elements from $\mathbb{F}_q$

Generator matrix in standard form (1)

Let C be an $[n, k]_q$ linear code. C has a unique generator matrix of the form

$$[I_k \mid \hat{G}].$$

A generator matrix in this form is said to be in standard form (or reduced echelon form).

Example (Binary Hamming code $n = 7$)

$$G = \left(I_4 \mid \hat{G} \right) = \left(\begin{array}{cccc|ccc} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{array} \right)$$

Systematic encoding: Encoding with a generator matrix in standard form.

$$(m_1 \dots m_k) \cdot \left(I_k \mid \hat{G} \right) = (m_1, \dots, m_k, *, \dots, *)$$

Generator matrix in standard form (2)

Corollary

Let C be an $[n,k]_d$ linear code, if $G = [I_k \mid \hat{G}]$ is a generator matrix in standard form, then $H = [-\hat{G}^T \mid I_{n-k}]$ is a parity-check matrix for C.

Proof.

Note that $\hat{G} \in \mathbb{F}_q^{k\times(n-k)}$ and that

$$G \cdot H^T = \left(I_k \,|\, \hat{G}\right) \cdot \begin{pmatrix} -\hat{G} \\ I_{n-k} \end{pmatrix} = -\hat{G} + \hat{G} = 0$$

Moreover H has $n-k$ linearly independent rows. This concludes the proof. □

Dual code

Since the $n-k$ rows of a parity-check matrix H are independent, H is a generator matrix too.

Definition

The *dual code* of C is the $[n, n-k]_q$ linear code $C^\perp$ composed by all the vectors orthogonal to all words of C:

$$C^\perp = \{\tilde{\mathbf{c}} \mid \tilde{\mathbf{c}} \cdot \mathbf{c} = 0, \forall \mathbf{c} \in C\}.$$

C	$C^\perp$
$[n, k]_q$ linear code	$[n, n-k]_q$ linear code
$G \in \mathbb{F}_q^{k \times n}$ generator matrix	$G \in \mathbb{F}_q^{k \times n}$ parity-check matrix
$H \in \mathbb{F}_q^{(n-k) \times n}$ parity-check matrix	$H \in \mathbb{F}_q^{(n-k) \times n}$ generator matrix

Distance of a linear code

What can we say about the distance of a linear code $[n, k, d]_q$?

Distance of a linear code

What can we say about the distance of a linear code $[n, k, d]_q$?

$$d = \min_{\mathbf{c} \in C, \mathbf{c} \neq 0} \mathrm{wt}(\mathbf{c}) = \mathrm{wt}(C)$$

Proof.

a. $d \leq \mathrm{wt}(C)$: this is trivial as $\mathbf{0} \in C$, so if $\mathbf{c} \in C$ is the codeword with minimum weight, we can compute $d(0, \mathbf{c}) = \mathrm{wt}(\mathbf{c})$.

b. $d \geq \mathrm{wt}(C)$: for any $\mathbf{c}_1 \neq \mathbf{c}_2 \in C$, we note that $\mathbf{c}_1 - \mathbf{c}_2 \in C$. Now note that the weight of $\mathbf{c}_1 - \mathbf{c}_2$ is $d(\mathbf{c}_1, \mathbf{c}_2)$ (why?), since the non-zero symbols in $\mathbf{c}_1 - \mathbf{c}_2$ occur exactly in the positions where the two codewords differ.

$\square$

We show the relation between the weight of a codeword and H

Theorem

If $\mathbf{c} \in C$, the columns of H corresponding to the nonzero coordinates of $\mathbf{c}$ are linearly dependent. Conversely, if a linear dependence relation with nonzero coefficients exists among w columns of H, then there is a codeword in C of weight w whose nonzero coordinates correspond to these columns.

Proof's idea: If for example $\mathrm{supp}(\mathbf{c}) = \{c_0, c_1, c_2\}$ then

$$0 = H\mathbf{c}^T = \begin{bmatrix} \mathbf{h}_0 & \mathbf{h}_2 & \dots & \mathbf{h}_{n-1} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \longrightarrow \mathbf{h}_0 c_0 + \mathbf{h}_1 c_1 + \mathbf{h}_2 c_2 = 0$$

If $\mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_3$ are linearly dependent, then exist $a_0, a_1, a_2 \in \mathbb{F}_q$ (not all zero) such that $a_0\mathbf{h}_0 + a_1\mathbf{h}_1 + a_2\mathbf{h}_2 = 0$.

We show the relation between the weight of a codeword and H

Theorem

If $\mathbf{c} \in C$, the columns of H corresponding to the nonzero coordinates of $\mathbf{c}$ are linearly dependent. Conversely, if a linear dependence relation with nonzero coefficients exists among w columns of H, then there is a codeword in C of weight w whose nonzero coordinates correspond to these columns.

For any $[n, k]_q$ code C with parity check matrix H, the distance $d(C)$ is such that

- $d(C) \geq d \iff$ every subset of $d-1$ columns of H are linearly independent
- $d(C) \leq d \iff$ there exists a subset of d columns of H that are linearly dependent

The main problem of coding theory

Consider an (n, M, d) code over an alphabet $\mathcal{A}$.

- The larger is the value M, the more efficient is the code

$$A_q(n, d) = \max\{M \mid \text{there exists an}(n, M, d)\text{-code over}\mathcal{A}\}$$

The main problem of coding theory

Consider an (n, M, d) code over an alphabet $\mathcal{A}$.

- The larger is the value M, the more efficient is the code

$$A_q(n,d) = \max\{M \mid \text{there exists an}(n,M,d)\text{-code over}\mathcal{A}\}$$

For practical purposes a "good" (n, M, d) code will have:

- small n
- large M (to permit a wide variety of messages);
- large d (for detecting and correcting large number of errors).

These are conflicting aims.
Thus we come to the *Main Problem of Coding Theory*:

Given a q-ary alphabet, a length n and a minimum distance d, find a code such that $A_q(n, d)$ is maximal.

Singleton bound

Theorem (Singleton Bound)

If C is an $(n, M, d)_q$ code, then $A_q(n, d) \leq q^{n-d+1}$

$$q^k \leq q^{n-d+1} \longrightarrow k \leq n - d + 1$$

Codes that meet this bound, i.e. satisfy $d = n - k + 1$, are called **Maximum Distance Separable** (MDS) codes.

Fix $n, k \in \mathbb{N}$, such $k \leq n$ and q a prime power with $q \geq n$. Consider the finite field $\mathbb{F}_q$ and construct the code as follows:

Fix $n, k \in \mathbb{N}$, such $k \leq n$ and q a prime power with $q \geq n$. Consider the finite field $\mathbb{F}_q$ and construct the code as follows:

1. Choose n distinct points $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$
2. Let $\mathbf{m} = (m_0, \ldots, m_{k-1})$ a message in $\mathbb{F}_q^k$, we can rewrite $\mathbf{m}$ as

$$m(x) = m_0 + m_1 x + \cdots + m_{k-1} x^{k-1} \in \mathbb{F}_q[x]$$

3. Encode $m(x)$ evaluating it in α_i, $i = 1, \ldots, n$:

$$c(x) = (m(\alpha_1), \ldots, m(\alpha_n)).$$

Definition (Reed-Solomon codes)

Take n distinct points $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$, with n such that $q \geq n$, and let k be an integer such that $1 \leq k \leq n$. We define the Reed-Solomon code as

$$RS_q(n, k) = \{(f(\alpha_1), \ldots, f(\alpha_n)) \in \mathbb{F}_q^n \mid f \in \mathbb{F}_q[x] s.t. \deg(f) \leq k-1 \cup \{0\}$$

Remark: Usually the set of points $S = \{\alpha_1, \ldots, \alpha_n\}$ is $\mathbb{F}_q^*$.

Definition (Reed-Solomon codes)

Take n distinct points $\alpha_1, \dots, \alpha_n \in \mathbb{F}_q$, with n such that $q \geq n$, and let k be an integer such that $1 \leq k \leq n$. We define the Reed-Solomon code as

$$RS_q(n,k) = \{(f(\alpha_1), \dots, f(\alpha_n)) \in \mathbb{F}_q^n \mid f \in \mathbb{F}_q[x] s.t. \deg(f) \leq k-1 \cup \{0\}$$

Remark: Usually the set of points $S = \{\alpha_1, \dots, \alpha_n\}$ is $\mathbb{F}_q^*$.

- Let $\mathcal{P}_{k-1}$ be the vector space of all polynomials of degree $k-1$ over $\mathbb{F}_q$

$$\{1, x, \dots, x^{k-1}\}$$

is a basis for it.

We can define a code $C = RS(n,k)$ as the image of

$$\begin{aligned} \mathrm{Enc} : \mathcal{P}_{k-1} &\longrightarrow \mathbb{F}_q^n \\ f \quad &\longmapsto (f(\alpha_1), \ldots, f(\alpha_n)) \end{aligned}$$

In this way the Vandermonde matrix

$$G = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \ldots & \alpha_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{k-1} & \alpha_2^{k-1} & \ldots & \alpha_n^{k-1} \end{pmatrix}$$

(obtained evaluating $\{1, x, \ldots, x^{k-1}\}$ in $\alpha_1, \ldots, \alpha_n$) is a generator matrix for C.

Example

Consider the RS codes over $\mathbb{F}_9$ with $k = 3$. Let $\{1, x, x^2\}$ a basis for $\mathcal{P}_2$. Then let S be the set of points $\mathbb{F}_9^* = \{1, \alpha, \alpha^2, \alpha^3, \alpha^4, \alpha^5, \alpha^6, \alpha^7\}$, we obtain the generator matrix

$$G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 & \alpha^7 \\ 1 & \alpha^2 & \alpha^4 & \alpha^6 & \alpha^1 & \alpha^2 & \alpha^4 & \alpha^6 \end{pmatrix}.$$

The corresponding Reed-Solomon code is a linear code with block length $n = q - 1 = 8$, and dimension $k = \dim \mathcal{P}_{k-1} = 3$.

Example

Consider the RS codes over $\mathbb{F}_9$ with $k = 3$. Let $\{1, x, x^2\}$ a basis for $\mathcal{P}_2$. Then let S be the set of points $\mathbb{F}_9^* = \{1, \alpha, \alpha^2, \alpha^3, \alpha^4, \alpha^5, \alpha^6, \alpha^7\}$, we obtain the generator matrix

$$G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 & \alpha^7 \\ 1 & \alpha^2 & \alpha^4 & \alpha^6 & \alpha^1 & \alpha^2 & \alpha^4 & \alpha^6 \end{pmatrix}.$$

The corresponding Reed-Solomon code is a linear code with block length $n = q - 1 = 8$, and dimension $k = \dim \mathcal{P}_{k-1} = 3$.

What about the distance?

Example

Consider the RS codes over $\mathbb{F}_9$ with $k = 3$. Let $\{1, x, x^2\}$ a basis for $\mathcal{P}_2$. Then let S be the set of points $\mathbb{F}_9^* = \{1, \alpha, \alpha^2, \alpha^3, \alpha^4, \alpha^5, \alpha^6, \alpha^7\}$, we obtain the generator matrix

$$G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 & \alpha^7 \\ 1 & \alpha^2 & \alpha^4 & \alpha^6 & \alpha^1 & \alpha^2 & \alpha^4 & \alpha^6 \end{pmatrix}.$$

The corresponding Reed-Solomon code is a linear code with block length $n = q - 1 = 8$, and dimension $k = \dim \mathcal{P}_{k-1} = 3$.

What about the distance?

The $RS_q(n, k)$ is MDS, i.e. it is an $[n, k, d = n - k + 1]_q$ code

Example

Consider the RS codes over $\mathbb{F}_9$ with $k = 3$. Let $\{1, x, x^2\}$ a basis for $\mathcal{P}_2$. Then let S be the set of points $\mathbb{F}_9^* = \{1, \alpha, \alpha^2, \alpha^3, \alpha^4, \alpha^5, \alpha^6, \alpha^7\}$, we obtain the generator matrix

$$G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 & \alpha^7 \\ 1 & \alpha^2 & \alpha^4 & \alpha^6 & \alpha^1 & \alpha^2 & \alpha^4 & \alpha^6 \end{pmatrix}.$$

The corresponding Reed-Solomon code is a linear code with block length $n = q - 1 = 8$, and dimension $k = \dim \mathcal{P}_{k-1} = 3$.

What about the distance?

The $RS_q(n, k)$ is MDS, i.e. it is an $[n, k, d = n - k + 1]_q$ code

The code above is an $[8, 3, 6]$ Reed-Solomon code over $\mathbb{F}_9$