Linear Algebra

Course 10: 10.12.2020

Chapter 4. Introduction to Coding Theory

Part I

Contents

Coding theory

Starting points:

- Shannon 1948: Information Theory
- Hamming 1950: Error-Correcting Codes

Main classes of codes:

- source coding: data compression
- channel coding: error-correcting codes

A first example

EAN-13 International Article Number

It is a sequence of 13 digits $a_1, a_2, \ldots, a_{13}$ that identifies a product. Digit a_{13} is a check digit that is computed as

$$a_{13} = 10 - (a_1 + 3a_2 + a_3 + 3a_4 + \cdots + a_{11} + 3a_{12}) \bmod 10.$$

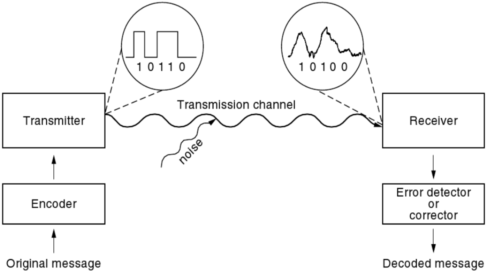
Digits are written in binary; black bars for 1, white bars for 0.

In particular:

- ISBN (International Standard Book Number)
- UPC (Universal Product Code) etc.

Error-correcting (detecting) codes

General scheme:

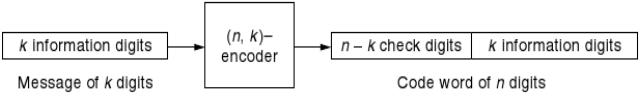


Different codes are suitable for different applications:

- satellite and space transmissions
- credit cards
- CD's, DVD's, Blu-ray discs etc.

The coding problem

- We discuss *binary codes*. In general: codes over finite fields.
- We consider *symmetric channels*: the probability of 1 being changed into 0 is the same as that of 0 being changed into 1.
- It is assumed that the number of errors is less than the number of correctly transmitted bits.
- We talk about (n, k)-*codes*:



There are 2^k possible messages, and so 2^k code words. There are 2^n possible words received.

Aim

Find the right balance between k and $n - k$.

Two simple codes - The $(3, 2)$-parity check code

- The check digit is the sum modulo 2 of the message digits.
- Encoding:

Message	Code word
00	000
01	101
10	110
11	011

How many errors can this code detect/correct?

- Decoding:

Received words	101	111	100	000	110
Parity check	passes	fails	fails	passes	passes
Decoded words	01	-	-	00	10

Two simple codes - The $(3,1)$-repeating code

- The two check digits repeat the message digit.
- Encoding:

Message	Code word
0	000
1	111

  How many errors can this code detect/correct?
- Decoding:

Received words	111	010	011	000
Decoded words	1	0	1	0

Hamming distance

Definition

The *Hamming distance* between two words of the same length is the number of positions in which they difer.

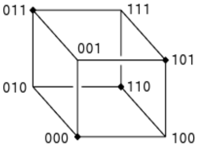
Notation $d(u, v)$.
Example: $d(101, 100) = 1$, $d(110, 001) = 3$, $d(101, 011) = 2$.

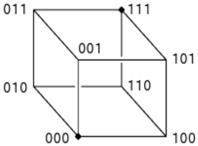
Theorem

The Hamming distance is a metric on the set $\mathbb{Z}_2^n$ of words of length n, that is, the following properties hold for every $u, v, w \in \mathbb{Z}_2^n$:
(1) $d(u, v) = d(v, u)$.
(2) $d(u, v) + d(v, w) \geq d(u, w)$.
(3) $d(u, v) \geq 0$ *with equality if and only if* $u = v$.

Hamming distance - cont.

- In an (n, k)-code, the 2^n received words can be thought of as placed at the vertices of an n-dimensional cube with unit sides.
- The Hamming distance between two words is the shortest distance between their corresponding vertices along the edges of the n-cube.
- The 2^k code words form a subset of the 2^n vertices, and the code has better error-correcting and error-detecting capabilities the farther apart these code words are.
- Cube representations of the $(3, 2)$-parity check and $(3, 1)$-repeating codes:





Error detection/correction capabilities

Theorem

A code detects all sets of t or fewer errors $\iff$ the minimum Hamming distance between code words is at least $t+1$.

Theorem

A code is capable of correcting all sets of t or fewer errors $\iff$ the minimum Hamming distance between code words is at least $2t+1$.

Code	Minimum distance between words	No. of detectable errors	No. of correctable errors	Information rate
(n,k)-code	d	d-1	$\leq \frac{d-1}{2}$	$\frac{k}{n}$
$(3,2)$-parity check code	2	1	0	$\frac{2}{3}$
$(3,1)$-repeating code	3	2	1	$\frac{1}{3}$

Polynomial representation

- A binary n-digit word $a_0 a_1 \dots a_{n-1}$ may be identified with a polynomial $a_0 + a_1 X + \cdots + a_{n-1} X^{n-1} \in \mathbb{Z}_2[X]$.

Definition

Let $p \in \mathbb{Z}_2[X]$ be of degree $n-k$. The *polynomial code generated by* p is an (n,k)-code whose code words are those polynomials of degree less than n which are divisible by p. Then the polynomial p is called the *generator* of the code.

- A message of length k is represented by a polynomial $m \in \mathbb{Z}_2[X]$ of degree less than k.
- Since the message is stored in the right hand side of a word, the message digits are carried by the higher-order coefficients of a polynomial. So we consider $m \cdot X^{n-k}$.

Polynomial representation - cont.

- To encode the message polynomial m we first use the Division Algorithm to find unique $q, r \in \mathbb{Z}_2[X]$ such that

$$m \cdot X^{n-k} = q \cdot p + r\,, \quad degree(r) < degree(p) = n - k.$$

Then the code polynomial is

$$v = r + m \cdot X^{n-k}.$$

The check digits of the message are carried by r.

Theorem

With the above notation, the code polynomial v is divisible by p.

Proof. We have $v = r + m \cdot X^{n-k} = r + q \cdot p + r = q \cdot p$, because $r \in \mathbb{Z}_2[X]$, and so $r + r = 0$.

Polynomial representation - examples

Example 1. *Let* $p = 1 + X^2 + X^3 + X^4 \in \mathbb{Z}_2[X]$ *be the generator polynomial of a* $(7,3)$*-code. Let us encode the message 101.*

Solution. Note that $n = 7$ and $k = 3$.

$$\begin{aligned}
\text{message } 101 &\rightsquigarrow m = 1 \cdot 1 + 0 \cdot X + 1 \cdot X^2 = 1 + X^2 \\
&\rightsquigarrow mX^{n-k} = (1 + X^2) \cdot X^4 = X^4 + X^6 \\
&\rightsquigarrow r = mX^{n-k} \bmod p = (X^4 + X^6) \bmod p = 1 + X \\
&\rightsquigarrow v = r + mX^{n-k} = 1 + X + X^4 + X^6 \\
&\rightsquigarrow \text{code word } \boxed{1100}\boxed{101}
\end{aligned}$$

Example 2. *If the generator polynomial of a* $(6,3)$*-code is* $p = 1 + X + X^3 \in \mathbb{Z}_2[X]$, *test whether the following received words contain detectable errors: 100011, 100110.*

Solution. We check if the received words are code words, that is, their associated polynomials are divisible by p [...].

Example 3. *Write down all the code words for the* $(6, 3)$*-code generated by the polynomial* $p = 1 + X + X^3 \in \mathbb{Z}_2[X]$.

Solution. Note that $n = 6$, $k = 3$, and we have $2^k = 8$ code words. We obtain the following table:

message	code word
000	000000
001	111001
010	011010
011	100011
100	110100
101	001101
110	101110
111	010111

E.g.: $110 \rightsquigarrow m = 1 + X \rightsquigarrow mX^{n-k} = X^3 + X^4$

$$\rightsquigarrow r = mX^{n-k} \bmod p = (X^3 + X^4) \bmod p = 1 + X^2$$

$$\rightsquigarrow v = r + mX^{n-k} = 1 + X^2 + X^3 + X^4 \rightsquigarrow \boxed{101}\boxed{110}$$

Matrix representation

- A binary n-digit word $a_0 a_1 \ldots a_{n-1}$ may be identified with a matrix $\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} \in M_{n,1}(\mathbb{Z}_2)$.
- For an (n, k)-code, we see the 2^k possible messages as the elements of the vector space $\mathbb{Z}_2^k$ over $\mathbb{Z}_2$, and the 2^n possible received words as the elements of the vector space $\mathbb{Z}_2^n$ over $\mathbb{Z}_2$.

Definition

- An *encoder* of an (n, k)-code is an injective function $\gamma : \mathbb{Z}_2^k \to \mathbb{Z}_2^n$ (or equivalently, $\gamma : M_{k,1}(\mathbb{Z}_2) \to M_{n,1}(\mathbb{Z}_2)$).
- An (n, k)-code is called *linear* if its encoder is a linear map.

From now on we will discuss only linear codes.
An example: *Reed-Solomon code*, used for CD's, DVD's, Blu-ray discs etc.

Theorem

Any (n, k)-code generated by a polynomial of degree $n - k$ is linear.

Proof. Let $p \in \mathbb{Z}_2[X]$ be the generator polynomial. We have seen that we encode the message $m \in \mathbb{Z}_2[X]$ as $v = r + m \cdot X^{n-k}$, where $r \in \mathbb{Z}_2[X]$ is the remainder of the division of m by p, that is, $m \bmod p$.

Hence the encoder $\gamma : \mathbb{Z}_2^k \to \mathbb{Z}_2^n$ associates to the k-tuple of the coefficients of m the n-tuple of the coefficients of $v = r + m \cdot X^{n-k}$.

One shows that γ is a linear map, that is,

$$\gamma(k_1 m_1 + k_2 m_2) = k_1 \gamma(m_1) + k_2 \gamma(m_2),$$

$\forall k_1, k_2 \in \mathbb{Z}_2, \forall m_1, m_2 \in \mathbb{Z}_2^k$.

Generator matrix

Definition

Consider a linear (n, k)-code with encoder $\gamma : \mathbb{Z}_2^k \to \mathbb{Z}_2^n$. Let E, E' be the canonical bases of the $\mathbb{Z}_2$-vector spaces $\mathbb{Z}_2^k$ and $\mathbb{Z}_2^n$ respectively. Then the matrix

$$G = [\gamma]_{EE'}$$

is called the *generator matrix* of the code.

A message $m \in \mathbb{Z}_2^k$ encodes as $\gamma(m)$.

But for $m \in \mathbb{Z}_2^k$, we have $[\gamma(m)]_{E'} = [\gamma]_{EE'} \cdot [m]_E$.

Hence a message $m \in M_{k,1}(\mathbb{Z}_2)$ encodes as $G \cdot m$.

Generator matrix - cont.

Use the above notation.

Theorem

(i) The code words of the (n,k)-code are the vectors in the subspace $\operatorname{Im}\gamma$ of $\mathbb{Z}_2^n$. Hence a binary (n,k)-code means a k-dimensional subspace of the vector space $\mathbb{Z}_2^n$.
(ii) The columns of G form a basis of this subspace, and so a vector is a code vector if and only if it is a unique linear combination of the columns of G.

Remark. A code word contains the message digits on the last k positions. Hence the generator matrix G of an (n,k)-code is always of the form

$$G = \begin{pmatrix} P \\ I_k \end{pmatrix} \in M_{n,k}(\mathbb{Z}_2),$$

where $P \in M_{n-k,k}(\mathbb{Z}_2)$ and $I_k \in M_k(\mathbb{Z}_2)$ is the identity matrix.

Parity check matrix

Definition

With the above notation, the matrix

$$H = \begin{pmatrix} I_{n-k} & P \end{pmatrix} \in M_{n-k,n}(\mathbb{Z}_2)$$

is called the *parity check matrix* of the code.

Theorem

Consider a linear (n, k)-code with parity check matrix $H = \begin{pmatrix} I_{n-k} & P \end{pmatrix} \in M_{n-k,n}(\mathbb{Z}_2)$. Then a received vector $u \in \mathbb{Z}_2^n$ (or $u \in M_{n,1}(\mathbb{Z}_2)$) is a code vector if and only if $H \cdot u = 0$.

Matrix representation - examples

Example 1. *Determine the generator matrix and the parity check matrix of the* $(3,2)$*-parity check code, and characterize the code vectors.*

Solution. Note that $n = 3$ and $k = 2$. The encoder is a $\mathbb{Z}_2$-linear map $\gamma : \mathbb{Z}_2^k \to \mathbb{Z}_2^n$, i.e. $\gamma : \mathbb{Z}_2^2 \to \mathbb{Z}_2^3$. The encoding of v is $\gamma(v)$.

- The generator matrix is $G = [\gamma]_{EE'}$, where E, E' are the canonical bases of $\mathbb{Z}_2^2$ and $\mathbb{Z}_2^3$ respectively.
  We have $e_1 = (1,0) \rightsquigarrow 10 \rightsquigarrow \boxed{1}\boxed{10} \rightsquigarrow (1,1,0) = \gamma(e_1)$.
  We have $e_2 = (0,1) \rightsquigarrow 01 \rightsquigarrow \boxed{1}\boxed{01} \rightsquigarrow (1,0,1) = \gamma(e_2)$.
  Hence $G = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} P \\ I_2 \end{pmatrix} = \begin{pmatrix} P \\ I_k \end{pmatrix}$.
- The parity check matrix is
  $H = \begin{pmatrix} I_{n-k} & P \end{pmatrix} = \begin{pmatrix} I_1 & P \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$.
- $(u_1, u_2, u_3) \in \mathbb{Z}_2^3$ is a code word $\Leftrightarrow H \cdot [u]_{E'} = [0]_{E'} \Leftrightarrow$ $u_1 + u_2 + u_3 = 0 \Leftrightarrow u_1 = u_2 + u_3$.

Example 2. *Determine the generator matrix and the parity check matrix of the* $(3,1)$*-repeating code, and characterize the code vectors.*

Solution. Note that $n = 3$ and $k = 1$. The encoder is a $\mathbb{Z}_2$-linear map $\gamma : \mathbb{Z}_2^k \to \mathbb{Z}_2^n$, i.e. $\gamma : \mathbb{Z}_2 \to \mathbb{Z}_2^3$. The encoding of v is $\gamma(v)$.

- The generator matrix is $G = [\gamma]_{EE'}$, where E, E' are the canonical bases of $\mathbb{Z}_2$ and $\mathbb{Z}_2^3$ respectively.
  We have $e_1 = 1 \rightsquigarrow \boxed{11}\boxed{1} \rightsquigarrow (1,1,1) = \gamma(e_1)$.
  Hence $G = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} P \\ I_1 \end{pmatrix} = \begin{pmatrix} P \\ I_k \end{pmatrix}$.
- The parity check matrix is
  $$H = \begin{pmatrix} I_{n-k} & P \end{pmatrix} = \begin{pmatrix} I_2 & P \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$
- $(u_1, u_2, u_3) \in \mathbb{Z}_2^3$ is a code word $\Leftrightarrow H \cdot [u]_{E'} = [0]_{E'} \Leftrightarrow$ $u_1 + u_3 = 0$ and $u_2 + u_3 = 0 \Leftrightarrow u_1 = u_2 = u_3$.

Example 3. *Determine the generator matrix and the parity check matrix of the* $(6,3)$*-code generated by the polynomial* $p = 1 + X + X^3 \in \mathbb{Z}_2[X]$*, and characterize the code vectors.*

Solution. Note that $n = 6$ and $k = 3$. The encoder is a $\mathbb{Z}_2$-linear map $\gamma : \mathbb{Z}_2^k \to \mathbb{Z}_2^n$, i.e. $\gamma : \mathbb{Z}_2^3 \to \mathbb{Z}_2^6$. The encoding of v is $\gamma(v)$.

- The generator matrix is $G = [\gamma]_{EE'}$, where E, E' are the canonical bases of $\mathbb{Z}_2$ and $\mathbb{Z}_2^3$ respectively. We have

$$
\begin{aligned}
e_1 = (1,0,0) \rightsquigarrow 100 \rightsquigarrow m = 1 \rightsquigarrow m \cdot X^{n-k} &= X^3 \\
&\rightsquigarrow r = m \cdot X^{n-k} \bmod p = X^3 \bmod p = 1 + X \\
&\rightsquigarrow v = r + m \cdot X^{n-k} = 1 + X + X^3 \\
&\rightsquigarrow \boxed{110}\boxed{100} \rightsquigarrow (1,1,0,1,0,0) = \gamma(e_1).
\end{aligned}
$$

Similarly, $e_2 = (0,1,0) \rightsquigarrow (0,1,1,0,1,0) = \gamma(e_2)$ and $e_3 = (0,0,1) \rightsquigarrow (1,1,1,0,0,1) = \gamma(e_3)$.

Matrix representation - examples

- Hence $G = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} P \\ I_3 \end{pmatrix} = \begin{pmatrix} P \\ I_k \end{pmatrix}$.
- The parity check matrix is
$$H = \begin{pmatrix} I_{n-k} & P \end{pmatrix} = \begin{pmatrix} I_3 & P \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$
- $(u_1, u_2, u_3, u_4, u_5, u_6) \in \mathbb{Z}_2^6$ is a code word $\Leftrightarrow H \cdot [u]_{E'} = [0]_{E'}$
$$\Leftrightarrow \begin{cases} u_1 + u_4 + u_6 = 0 \\ u_2 + u_4 + u_5 + u_6 = 0 \\ u_3 + u_5 + u_6 = 0 \end{cases} \Leftrightarrow \begin{cases} u_1 = u_4 + u_6 \\ u_2 = u_4 + u_5 + u_6 \\ u_3 = u_5 + u_6 \end{cases}.$$